

Helpin, Luke

378.744

BD

AM 1933

na



Boston University  
College of Liberal Arts  
Library

THE GIFT OF The Author

378.744

BD

AM 1933

ha

p7493



BOSTON UNIVERSITY GRADUATE SCHOOL

Thesis

Simple Correlation Including the Correlation  
Coefficient, Correlation from Ranks, and Mean  
Square Contingency.

by

Luke Halpin (A.B. Bowdoin 1921)

Submitted in partial fulfilment of the requirements  
for the degree of Master of Arts.

1933

BOSTON UNIVERSITY  
COLLEGE OF LIBERAL ARTS  
LIBRARY

HAL

77493



378.744

BO

AM 1933

ha

## OUTLINE

### Chapter

- I. A General Discussion of the Problem of Correlation
- II. Definition of Correlation and an Introduction to the Regression Method
- III. Development of the Regression Lines
  - A. An Indirect Geometric Approach
  - B. An Analytic Approach
  - C. A Development Showing the Range of Values of  $r$ .  
An Exact Definition of Correlation
  - D. The Standard Deviation of the Rows and Columns  
from the Regression Lines
- IV. The Correlation Surface
- V. The Product-Moment Formula for Correlation
- VI. Some Interesting Points Arrived at by Considering  
Normal Correlation
- VII. Computation Formulas for the Coefficient of Correlation.  
Problem
- VIII. Correlation from Ranks
- IX. Mean Square Contingency



# CONTENTS

|     |  |
|-----|--|
| 1.  | A general statement of the theories of correlation                           |
| 2.  | Definition of correlation and its relation to the regression method          |
| 3.  | Calculation of the regression lines  |
| 4.  | A. An indirect method of regression  |
| 5.  | B. An algebraic method   |
| 6.  | C. Determination of the value of $r$ by the method of least squares          |
| 7.  | D. The standard deviation of the error and its effect on the regression line |
| 8.  | The correlation surface  |
| 9.  | The product-moment formula for correlation                                   |
| 10. | Some properties of the partial coefficient of correlation                    |
| 11. | Correlation formulae for the coefficient of correlation                      |
| 12. | Correlation from ranks   |
| 13. | Some remarks on correlation  |



## Chapter I

A General Discussion of CorrelationA. A General Definition of Correlation

When a series of measures in statistics is studied, it is, in general, desirable to determine three points: (1) the computation of some average, the mean, median, or mode to represent the series; (2) the picturing of the degree of concentration by obtaining a measure of the dispersion; (3) the graphic picture of the distribution by plotting the smoothed frequency curve. When two or more series are to be compared, it is often necessary to find some method of determining the relationship between the series. This relationship is called correlation.

Suppose we consider a hypothetical case in discussing the correlation between marks given to a class of twenty pupils in plane geometry and English.

## School Marks Given a Class of Twenty Pupils in Plane Geometry and English

| Pupils | Average Marks in<br>Plane Geometry | Average Marks<br>in English | Rank in<br>Achievement in<br>Plane Geometry | Rank in<br>Achievement in<br>English |
|--------|------------------------------------|-----------------------------|---|--------------------------------------|
| A      | 50                                 | 60                          | 20  | 18                                   |
| B      | 68                                 | 72                          | 17  | 14                                   |
| C      | 92                                 | 85                          | 4   | 8                                    |
| D      | 84                                 | 91                          | 7   | 5                                    |
| E      | 97                                 | 96                          | 2   | 2                                    |
| F      | 72                                 | 80                          | 15  | 10                                   |
| G      | 82                                 | 75                          | 9   | 12                                   |
| H      | 76                                 | 84                          | 13  | 9                                    |
| I      | 62                                 | 50                          | 18  | 20                                   |
| J      | 56                                 | 55                          | 19  | 19                                   |
| K      | 85                                 | 90                          | 6   | 6                                    |
| L      | 98                                 | 97                          | 1   | 1                                    |
| M      | 90                                 | 79                          | 5   | 11                                   |
| N      | 70                                 | 61                          | 16  | 17                                   |
| O      | 83                                 | 92                          | 8   | 4                                    |
| P      | 80                                 | 74                          | 11  | 13                                   |
| Q      | 96                                 | 95                          | 3   | 3                                    |
| R      | 75                                 | 71                          | 14  | 15                                   |
| S      | 81                                 | 86                          | 10  | 7                                    |
| T      | 78                                 | 70                          | 12  | 16                                   |



A General Method of Correlation

When a series of measures in statistics is studied, it is, in general,

possible to determine three values: (1) the magnitude of each series,

the mean, median, or mode in respect to the series; (2) the nature of

the degree of concentration by obtaining a measure of the dispersion;

(3) the graphic character of the distribution by plotting the associated

frequency curve. When two or more series are to be compared, it is often

necessary to find some method of determining the relationship between the

series. This relationship is called correlation.

Suppose we consider a hypothetical case in illustrating the correlation

between words given to a class of twenty pupils in their geometry and English.

Suppose that there is a class of twenty pupils in plane geometry and English

| Pupils | Geometry Marks in<br>Plane Geometry | Geometry Marks in<br>English | Rank in<br>Geometry | Rank in<br>English |
|--------|-------------------------------------|------------------------------|---------------------|--------------------|
| 1      | 80                                  | 80                           | 1                   | 1                  |
| 2      | 75                                  | 75                           | 2                   | 2                  |
| 3      | 70                                  | 70                           | 3                   | 3                  |
| 4      | 65                                  | 65                           | 4                   | 4                  |
| 5      | 60                                  | 60                           | 5                   | 5                  |
| 6      | 55                                  | 55                           | 6                   | 6                  |
| 7      | 50                                  | 50                           | 7                   | 7                  |
| 8      | 45                                  | 45                           | 8                   | 8                  |
| 9      | 40                                  | 40                           | 9                   | 9                  |
| 10     | 35                                  | 35                           | 10                  | 10                 |
| 11     | 30                                  | 30                           | 11                  | 11                 |
| 12     | 25                                  | 25                           | 12                  | 12                 |
| 13     | 20                                  | 20                           | 13                  | 13                 |
| 14     | 15                                  | 15                           | 14                  | 14                 |
| 15     | 10                                  | 10                           | 15                  | 15                 |
| 16     | 5                                   | 5                            | 16                  | 16                 |
| 17     | 0                                   | 0                            | 17                  | 17                 |
| 18     | 80                                  | 80                           | 18                  | 18                 |
| 19     | 75                                  | 75                           | 19                  | 19                 |
| 20     | 70                                  | 70                           | 20                  | 20                 |



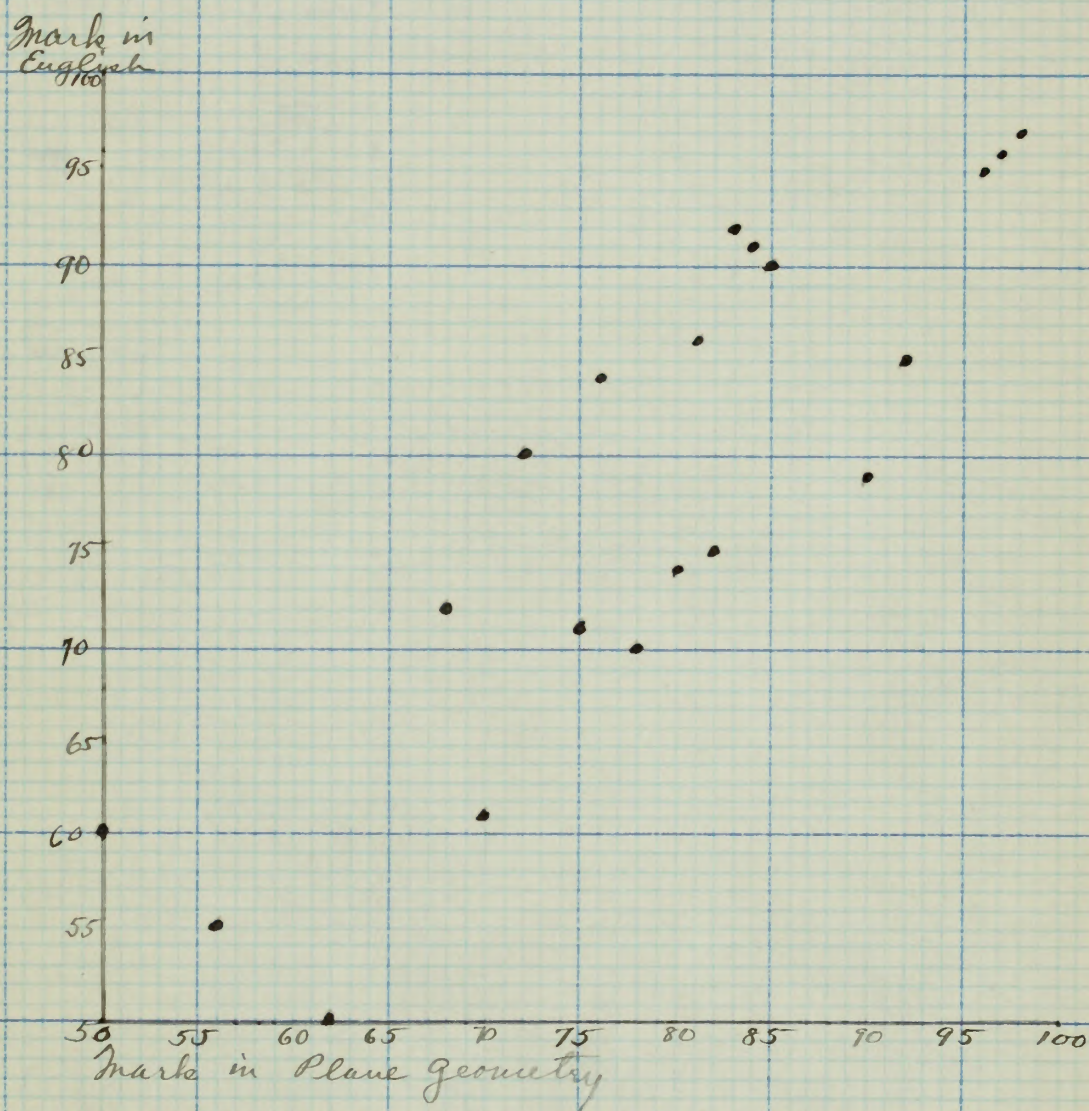
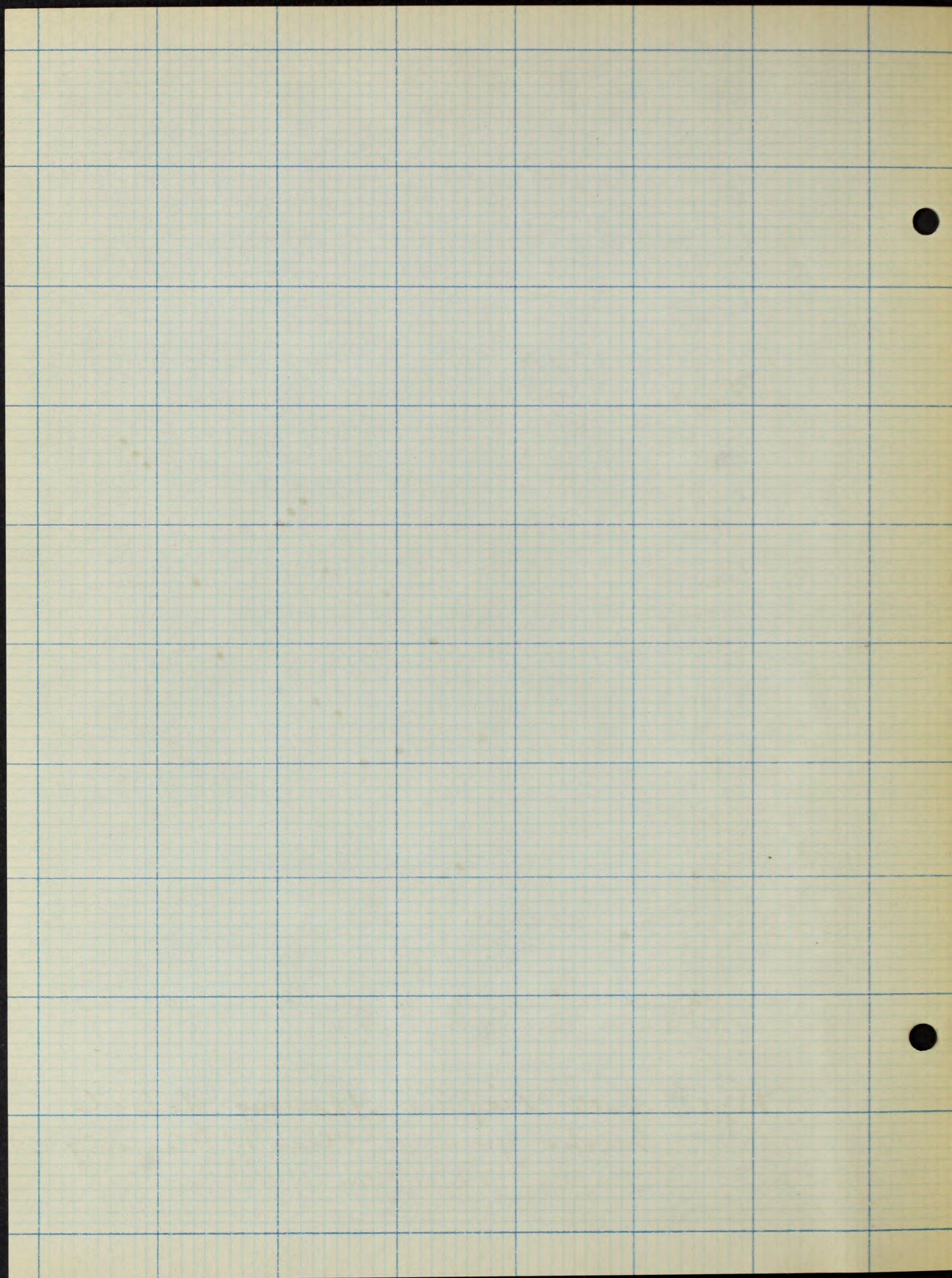


Fig I. Scatter Diagram Showing School Marks Given a Class of Twenty Pupils in Plane Geometry and English





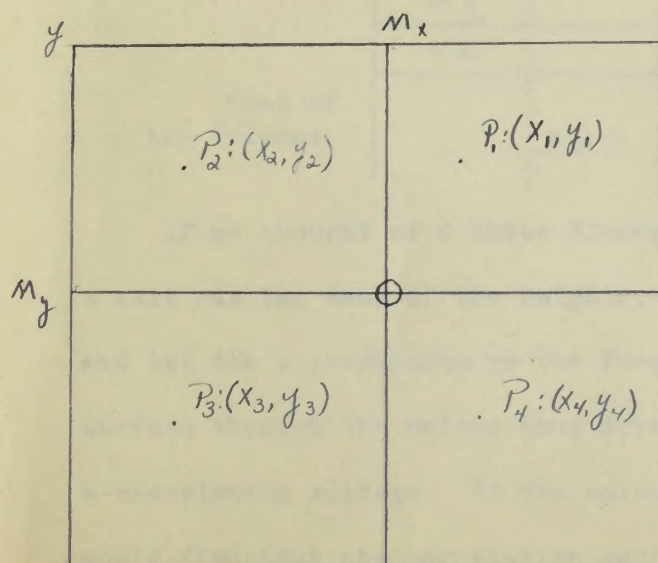


The graph would show readily that a pupil who stood high in plane geometry would also receive high marks in English and if low in geometry was also low in English. This graphic method would give a fair idea of the relationship existing between the two series but would not give an exact numerical expression nor would it give an expression which would summarize the situation.

The problem, then, is to find some device which would yield a numerical expression that would completely describe the relation existing between the two series. This numerical expression is called the "coefficient of correlation." The coefficient of correlation is the term used generally in statistics to refer to the one obtained by the product-moment method and is designated by "r". It is an index of linear correlation which type will be discussed in this paper.

The series in Fig. 1 form an example of linear correlation because the points tend to form a straight band across the graph. If this were perfect linear correlation, the points would lie on a straight line. Thus we might define correlation as the "tendency for two observed variables to be related in the form of a single-valued mathematical function."

The product-moment formula will be developed later, but a brief explanation of it may aid in stating what is meant by correlation. A measure



of relationship between the variables might be obtained by considering products of the deviations from the means expressed in terms of the standard deviations. Let  $x = \bar{X} - M_x$  and  $y = \bar{Y} - M_y$ ,  $\sigma_x = \frac{1}{N} \sum x^2$  and  $\sigma_y = \frac{1}{N} \sum y^2$ , then we could sum the  $N$  pairs of  $x$  products:

$$x \frac{x_1}{\sigma_x} \cdot \frac{y_1}{\sigma_y} + \frac{x_2}{\sigma_x} \cdot \frac{y_2}{\sigma_y} + \dots + \frac{x_n}{\sigma_x} \cdot \frac{y_n}{\sigma_y}$$



The first would show results that a small error in the  
 second would also result in a small error in the third and so on. It is  
 also in this way. This method would give a fair idea of the re-  
 lationship existing between the two series but would not give an exact  
 numerical expression nor would it give an expression which would be  
 the same.

The problem, then, is to find some device which would yield a numerical  
 expression that would completely describe the relation existing between the  
 two series. This numerical expression is called the "coefficient of corre-  
 lation". The coefficient of correlation is the term used generally in  
 statistics to refer to the one obtained by the product-moment method and is  
 denoted by  $r$ . It is an index of linear correlation which will be  
 discussed in this paper.

The article in this form is an example of linear correlation because the  
 points tend to form a straight line across the graph. If this were perfect  
 linear correlation, the points would lie on a straight line. Thus we might  
 expect correlation as the "measure" for two observed variables to be related  
 in the form of a single-valued mathematical function.

The product-moment formula will be developed later, but a brief ex-  
 planation of it may aid in stating what is meant by correlation. A measure  
 of relationship between the variables  
 might be obtained by considering each  
 one of the deviations from the mean  
 expressed in terms of the standard  
 deviations. Let  $X = \Delta_1, \Delta_2, \Delta_3, \dots, \Delta_n$  and  
 $Y = \delta_1, \delta_2, \delta_3, \dots, \delta_n$  and  
 then we could say the  $n$  pairs of  $X$   
 and  $Y$  are  

$$\frac{X_1}{\sigma_X} \cdot \frac{Y_1}{\sigma_Y}, \frac{X_2}{\sigma_X} \cdot \frac{Y_2}{\sigma_Y}, \frac{X_3}{\sigma_X} \cdot \frac{Y_3}{\sigma_Y}, \dots, \frac{X_n}{\sigma_X} \cdot \frac{Y_n}{\sigma_Y}$$
 and the average of these products is the coefficient of correlation  $r$ .

and divide by  $\frac{N}{N}$ . The result  $\frac{\sum xy}{N \sigma_x \sigma_y}$  is presented by  $r$  and is called the product-moment formula for correlation. It will be shown later that  $r$  may vary from  $-1$  (perfect negative correlation) thru  $0$  (lack of correlation) to  $+1$  (perfect positive correlation).

#### B. The Correlation Table and Correlation Surface.

If we were to consider the following problem to find the coefficient of correlation between the two series,

(1) Heights in inches of Glasgow school boys, ages 4.5 to 5.5 years,

and (2) Weights in pounds of these same boys,

the work would be arranged in a double entry table called a correlation table. In this table the frequencies are thought of as being concentrated at the midpoint of the class intervals; that is, the weights are divided into class intervals as follows:

24-28, 28-33, 34-38, etc., with 26, 31, 36, etc., the midpoints.

|  | Height<br>in Inches | Weight in Pounds |      |      |      |      |      |
|--|---------------------|------------------|------|------|------|------|------|
|  |                     | 26               | 31   | 36   | 41   | 46   | 51   |
| Forsyth<br>"Mathematical<br>Analysis of<br>Statistics"<br>P. 219 | 31                  | 2                |      |      |      |      |      |
|  | 34                  | 5                | 15   | 5    |      |      |      |
|  | 37                  | 1                | 18   | 72   | 8    |      |      |
|  | 40                  |                  | 5    | 87   | 90   | 7    | 1    |
|  | 43                  |                  |      | 4    | 35   | 21   | 5    |
|  | 46                  |                  |      | 1    |      | 2    |      |
| Mean of<br>the Columns   |                     | 33.7             | 36.2 | 38.7 | 40.6 | 42.5 | 43.5 |

If we thought of a three dimensional coordinate system in which the  $x$  axis was the mean of the heights, the  $y$  axis the mean of the weights, and let the  $z$  coordinates be the frequencies, and if we passed a smooth surface through the points thus determined, we would get what is known as a correlation surface. If the correlation table were symmetrical, we would find that the correlation surface was a normal surface; that is, a







bell-shaped surface with the  $z$  axis the centroid vertical.

### C. Methods of Approach to the Problem of Correlation.

Rietz "Mathematical Statistics" P. 77

There are two methods of approach to the problem of correlation: one is the "regression" method, the other the "correlation surface" method.

#### The Regression Method.

If we consider associated values of  $x$  and  $y$  as plotted in a scatter diagram and separate the dots into classes by selecting class intervals  $dx$  and  $dy$ , the  $y$ 's corresponding to any class  $dx$  are called an array of  $y$ 's and similarly the values of  $x$  corresponding to any interval  $dy$  are called an  $x$ -array. The regression curve  $y = f(x)$  is defined as the locus of the expected value of  $y$  in the array which corresponds to an assigned value of  $x$  as  $dx$  approaches zero; that is, the regression curve of  $y$  on  $x$  is the locus of the means of the arrays of  $y$ 's as  $dx$  approaches zero. Similarly the regression curve of  $x$  on  $y$  is the locus of the means of the arrays of  $x$ 's as  $dy$  approaches zero. Having found the regression curves of  $y$  on  $x$  and  $x$  on  $y$ , we are now interested in the distribution of the values of  $y$  whose average we have predicted. This is accomplished by measuring the dispersion of the values of  $y$  which correspond to an assigned value of  $x$ . In other words, we wish to know the average standard deviation of a row about the line which represents the locus of the means of the rows and also the average standard deviation of a column about the line which represents the locus of the means of the  $x$ -arrays.

To illustrate the regression method we might consider a problem of correlating the marks of a class in geometry and of the same class in English. We would first find a means of predicting the mean mark of a sub-group in the geometry class which had received identical marks in English, then we would find a measure to predict the dispersion of such a subgroup.



Self-correlation coefficient with the same time series.

## 2. Methods of approach to the problem of correlation.

There are two methods of approach to the problem of correlation:

1. The "regression" method, in which the "regression surface" is found.

The regression method.

If we consider a set of values of  $x$  and  $y$  as obtained in a series

of observations, and separate the data into classes by selecting class intervals

for  $x$  and  $y$ , the  $y$ 's corresponding to any class  $x$  are called an array of

$y$ 's and similarly the values of  $x$  corresponding to any interval of  $y$  are

called an array of  $x$ 's. The regression curve  $y = f(x)$  is defined as the locus

of the expected value of  $y$  in the array which corresponds to an assigned

value of  $x$  as the argument. That is, the regression curve of  $y$  on  $x$

is the locus of the means of the arrays of  $y$ 's as the argument varies.

Similarly the regression curve of  $x$  on  $y$  is the locus of the means of the

arrays of  $x$ 's as the argument varies. Having found the regression curves

of  $y$  on  $x$  and  $x$  on  $y$ , we are now interested in the distribution of the

values of  $y$  when  $x$  is given. This is accomplished by

considering the distribution of the values of  $y$  which correspond to an assigned

value of  $x$ . In other words, we wish to know the average standard deviation

of  $y$  when  $x$  is given. This is accomplished by the locus of the means of the

arrays of  $y$ 's when  $x$  is given. This is accomplished by the locus of the

means of the arrays of  $y$ 's when  $x$  is given.

To illustrate the regression method we shall consider a problem of

correlating the marks of a class in geometry and in English in 1921.

We would first find a means of predicting the mean mark of a sub-group in

geometry when which had received identical marks in English. Then we could

find a means to predict the distribution of such a sub-group.



### The Correlation Surface Method

In this method, we attempt to determine the probability,  $\phi(x,y)dx dy$ , that a pair of associated values  $(x,y)$  of  $x$  and  $y$  will fall into the rectangular area bounded by  $(x+dx)$  and  $(y+dy)$ . If  $m(x)$  is such that  $m(x)dx$  gives, to within infinitesimals of higher order, the probability that any  $x$  lies between  $x$  and  $(x+dx)$  and  $n(x,y)dy$  gives the probability that any  $y$  taken from the array which corresponds to the  $x$  chosen above will lie between  $y$  and  $(y+dy)$  then the probability that both will happen is

$$\phi(x,y) dx dy = m(x) n(x,y) dx dy.$$

We are thus able to set up the equation for the frequency surface  $z = \phi(x,y)$  and by a study of this to arrive at the coefficient of correlation between  $x$  and  $y$ .



In this report, an attempt is made to determine the probability that a pair of associated values ( $x, y$ ) of  $x$  and  $y$  will fall into the range  $x_1$  to  $x_2$  and  $y_1$  to  $y_2$ . It is assumed that the joint probability  $P(x, y)$  is a function of  $x$  and  $y$  and that the marginal probabilities  $P(x)$  and  $P(y)$  are known. The probability that any  $x$  lies between  $x_1$  and  $x_2$  is  $P(x_1, x_2)$  and the probability that any  $y$  lies between  $y_1$  and  $y_2$  is  $P(y_1, y_2)$ . The probability that both  $x$  and  $y$  lie in these ranges is  $P(x_1, x_2, y_1, y_2)$ .

It is assumed that the joint probability  $P(x, y)$  is a function of  $x$  and  $y$  and that the marginal probabilities  $P(x)$  and  $P(y)$  are known. The probability that both  $x$  and  $y$  lie in these ranges is  $P(x_1, x_2, y_1, y_2)$ .



## Chapter II

A Definition of Correlation and an Introduction to  
the Regression Method.

Having given a general idea of the problem we might now define correlation and then indicate how we would approach a solution by way of the regression method.

Definition: (Prof. Dow in a course in Statistics) "A quantity is said to be correlated with another quantity if to any value of the one quantity there exists a probable value of the other quantity and more exactly we shall call  $x$  and  $y$  correlated if when any particular values of  $y$  are selected, the average value of the corresponding  $x$  is thereby determined."

If we consider a problem like that given on page 4 and set up a graph in which we plotted only the mean value of the heights corresponding to any given weight, we would have a graph like the one pictured on page 9. The dots represent the mean values of the heights corresponding to the actual values of the weights. The line  $O, M_y$  cuts the  $Y$  axis at the mean of the heights and  $O, M_x$  cuts the  $X$  axis at the mean of the weights. The line  $CC'$  is fitted by inspection as being the line of the best fit which corresponds to the actual line of the means. This line serves the purpose of a generalized trend of the points. Since  $CC'$  is the line of best fit of the means of the columns, it must pass through  $O$ , the mean point of the entire distribution and if  $B; (x, y)$ , a point on the line, is taken so that  $x$  and  $y$  represent the deviations of this point from the means  $M_x$  and  $M_y$ , then the slope of this line is  $\frac{y}{x}$  or is the deviation of the point from the mean of the  $Y$ 's divided by the deviation of the point from the mean of the  $X$ 's. Since the slope is always the same and since the line passes through  $O$ , we may consider this the origin and write the equation of  $CC'$ :  $y = mx$ . Now if we find by measurement the  $x$  and  $y$  value of any one point, we may find



...the problem of the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...

...the ...  
...the ...  
...the ...



m and thus write the equation of the line  $CC'$ . The difficulty here is that y is measured in inches and x is measured in pounds. This may be cared for by dividing each by its standard deviation which measures the variation of each series about its mean. Therefore, if we consider the ratio  $\frac{y}{\sigma_y} \div \frac{x}{\sigma_x}$  we have a measure of the degree of relationship showing the trend of the variations of the x's and y's. This ratio  $\frac{\frac{y}{\sigma_y}}{\frac{x}{\sigma_x}}$  we define as r, the coefficient of correlation.

Therefore, we write  $\frac{y}{\sigma_y} = r \frac{x}{\sigma_x}$  as the equation of the line  $CC'$ . Similarly if we found the line best fitting the means of the rows and called it  $RR'$ , we could show the equation of  $RR'$  to be  $\frac{x}{\sigma_x} = r \frac{y}{\sigma_y}$  in the same way.

Having explained the meaning of correlation in terms of these equations, we shall now attempt to develop them in a more rigorous manner.



a first trial the coefficient of the line  $0.2$ . The  $100\%$  point is  $100$ .  
 It is measured in inches and  $x$  is measured in minutes. This can be used for  
 by dividing each by its standard deviation which measures the variation of  
 each series about its mean. Therefore, if we consider the ratio  

$$\frac{y}{\sigma_y} = \frac{x}{\sigma_x}$$
  
 we have a measure of the degree of relationship showing the trend of the  
 variation of the  $x$ 's and  $y$ 's. This ratio we define as  $r$ , the  
 coefficient of correlation.

Therefore, we write  $r = \frac{y}{\sigma_y} = \frac{x}{\sigma_x}$  as the equation of the line  $0.2$ .  
 Similarly if we found the line best fitting the means of the rows and called  
 it  $0.5$ , we would have the equation of  $0.5$  for  $r$ .  
 In the same way.

Having explained the meaning of correlation in terms of these equations,  
 we shall now attempt to develop them in a more rigorous manner.



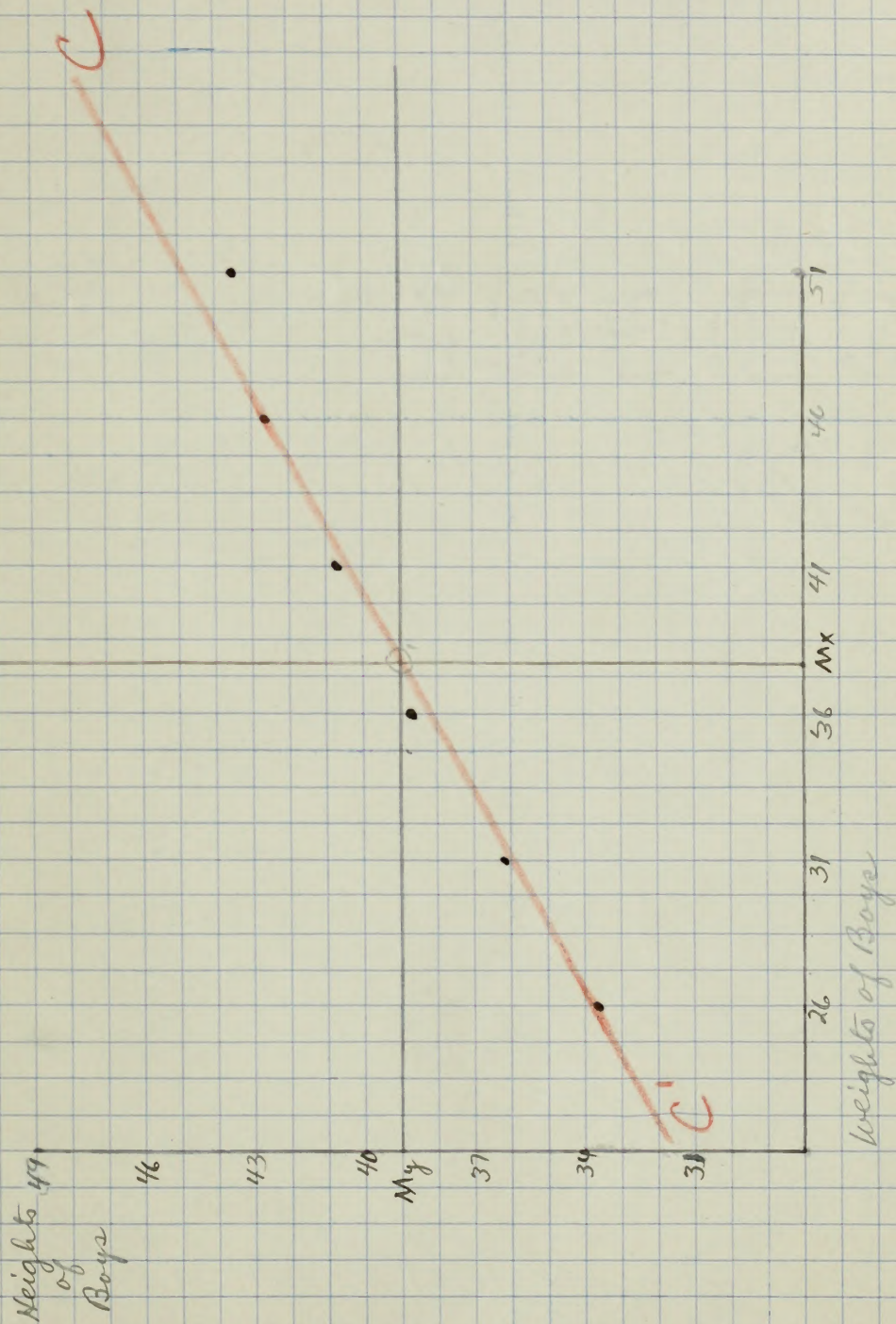
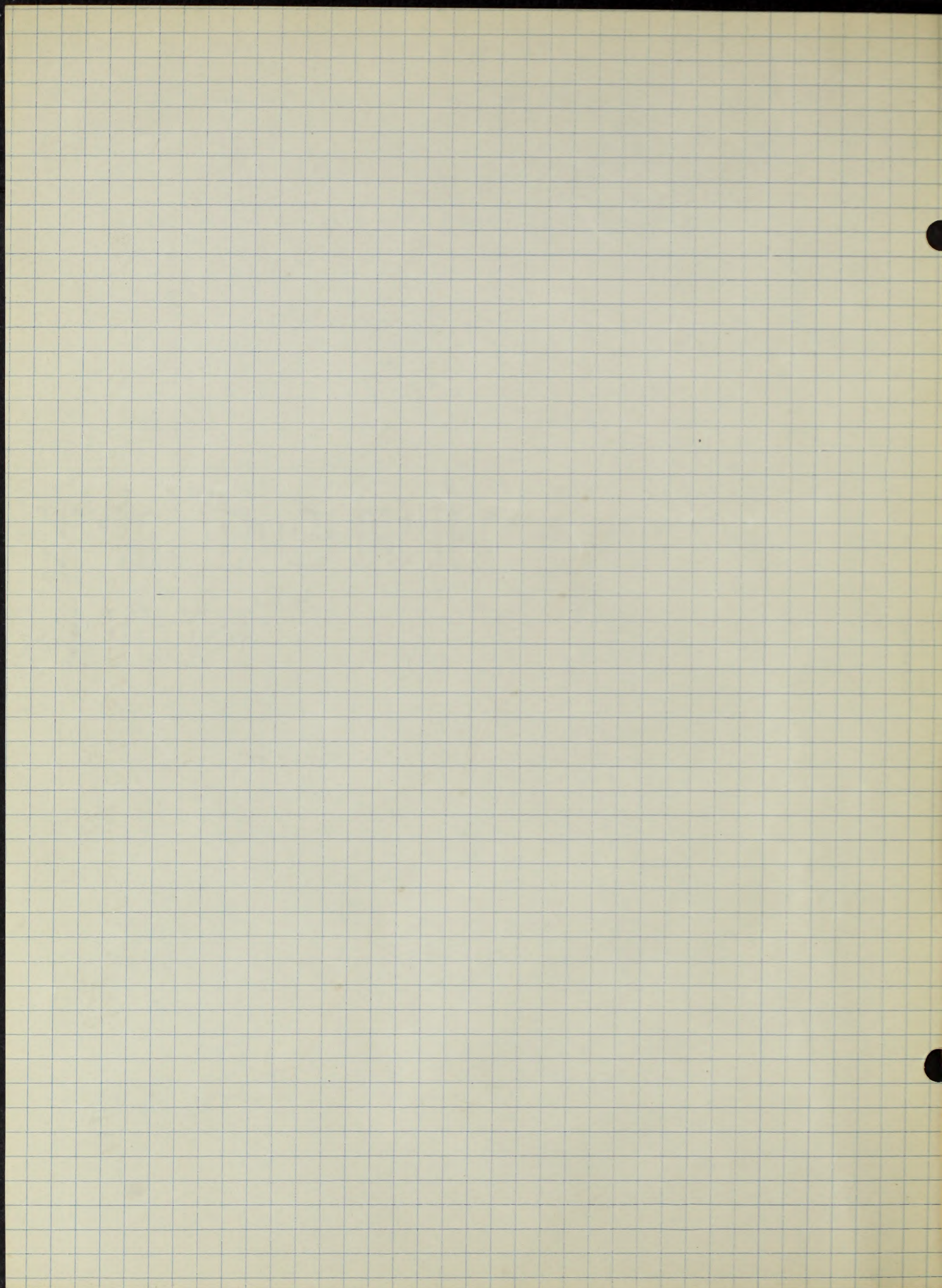


Fig. II. Graph Showing Estimated Line of Best Fit for the Means of the Columns.  
 $C'$  is the Regression Line of  $y$  on  $x$



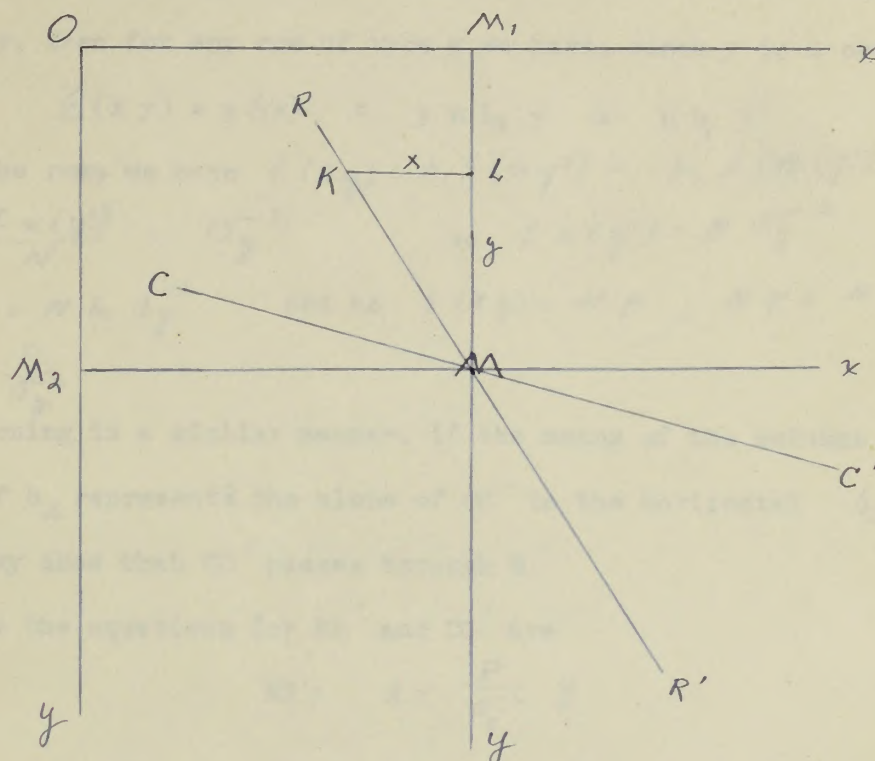




## Chapter III

Development of the Regression Lines.A. An Indirect Geometrical Approach

"Introduction to Theory of Statistics" Yule,  
P. 170 ff.



Suppose we had a distribution in which all the means of the rows were on the line  $RR'$  and let  $x$  be the deviations of the rows from  $M_1y$  and  $y$  be the deviations from  $M_2x$ . Call the slope of the line  $RR'$  to  $M_1y$ ,  $b$ ; the equation of  $RR'$  is then  $x = b, y$ . Then in any row of type  $y$  in which the number of observations is  $n$ ,  $\frac{\sum(x)}{n}$  = the deviation of the mean point of that row. Since that point is on  $RR'$ , we may now rewrite the equation of  $RR'$ :

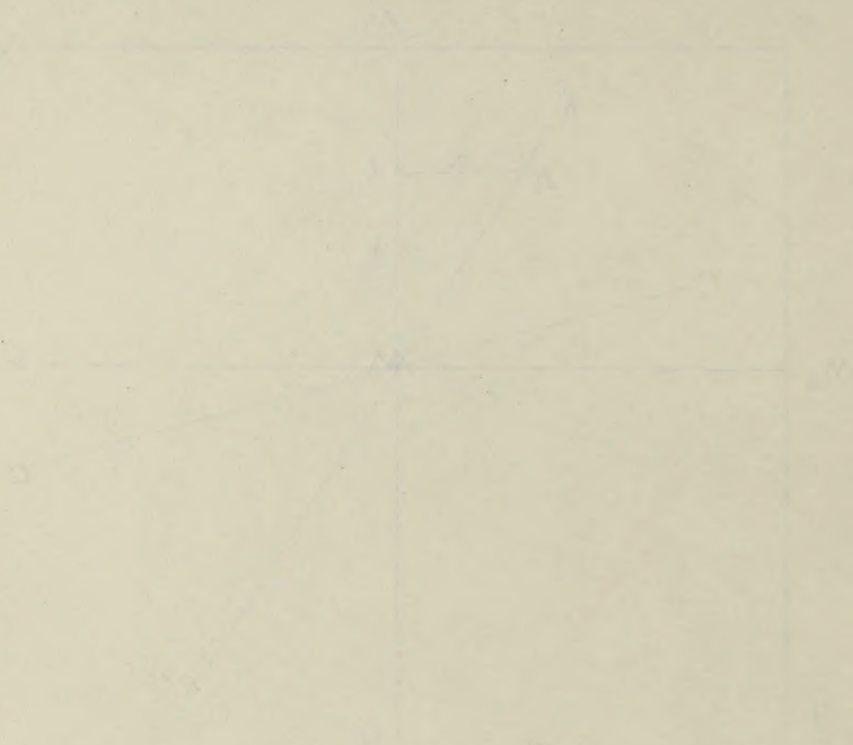
$$\frac{\sum(x)}{n} = b, y \quad \text{or} \quad \sum(x) = n b, y$$

If we consider this for the entire distribution, we write  $\sum x = b, \sum(ny)$  where  $\sum(x)$  is the sum of the deviations of all the  $X$ 's from  $M_1y$  and  $\sum(ny)$  is the sum of the deviations of all the  $\bar{y}$ 's from  $M_2x$ . But  $\sum(ny) = 0$  because  $M_2x$  is the mean of all the  $\bar{y}$ 's.  $\therefore \sum(x) = b, \sum(ny) = 0$

Since  $\sum(x) = 0$ , the sum of the deviations of all the  $\bar{X}$ 's from  $M_1y$  is zero, so  $M_1y$  must be the mean of all the  $\bar{X}$ 's and must cut  $O\bar{X}$  at  $M_1$ , the mean of  $\bar{X}$ . In this way we have shown  $M$  to be the mean of the entire distribution.



2. Introduction to the Theory of Statistics  
 Introduction to the Theory of Statistics



Suppose we have a distribution in which all the means of the rows were on the line  $x = \bar{x}$ , and let  $y$  be the deviation of the row from  $\bar{x}$  and  $y$  be the deviation from  $\bar{x}$ . Call the slope of the line  $\bar{x}$  to  $\bar{y}$  the correlation of  $\bar{x}$  is then  $x = \bar{x} + y$ . Then in any row of type  $y$  in which the number of observations is  $n$ , the deviation of the mean point of that row, since that point is on  $\bar{x}$ , we may now write the deviation of  $\bar{x}$  as  $\bar{x} - \bar{x} = 0$  or  $\bar{x} - \bar{x} = 0$ .

If we consider this for the entire distribution, we write  $\bar{x} - \bar{x} = 0$  or  $\bar{x} - \bar{x} = 0$ .

At this point is the sum of the deviations of all the  $\bar{x}$ 's from the mean  $\bar{x}$  is the sum of the deviations of all the  $\bar{x}$ 's from  $\bar{x}$ . But  $\bar{x} - \bar{x} = 0$  because  $\bar{x}$  is the mean of all the  $\bar{x}$ 's.

Since  $\bar{x} - \bar{x} = 0$ , the sum of the deviations of all the  $\bar{x}$ 's from  $\bar{x}$  is zero.

So  $\bar{x}$  must be the mean of all the  $\bar{x}$ 's and must satisfy  $\bar{x} - \bar{x} = 0$ , the mean of  $\bar{x}$ .

In this way we have shown  $\bar{x}$  to be the mean of the entire distribution.

Now  $RR'$  passes through M, a point which we may locate for any distribution. Therefore, to write the equation for the line we have only to determine  $b_1$ , its slope.

If we define  $p = \frac{1}{N} \sum (x, y)$  the mean product of all associated deviations of  $x$  and  $y$ , then for any row of type  $y$  we have, since  $y$  is a constant

$$\sum (x y) = y \sum (x) = y n b_1 y = n b_1 y^2$$

For all the rows we have  $\sum (x y) = b_1 \sum (n y^2) = b_1 \sum (n) (y^2)$

but  $\frac{\sum n (y^2)}{N} = \sigma_y^2$  so  $\sum n (y^2) = N \sigma_y^2$

$\therefore \sum (x y) = N b_1 \sigma_y^2$  and as  $\sum (x y) = N p$ ,  $N p = N b_1 \sigma_y^2$

and  $b_1 = \frac{p}{\sigma_y^2}$

Reasoning in a similar manner, if the means of the columns all lie on  $CC'$  and if  $b_2$  represents the slope of  $CC'$  to the horizontal  $b_2 = \frac{p}{\sigma_x^2}$ . Also we may show that  $CC'$  passes through M.

Hence the equations for  $RR'$  and  $CC'$  are

$$RR': \quad x = \frac{p}{\sigma_y^2} y$$

$$CC': \quad y = \frac{p}{\sigma_x^2} x$$

The forms of these equations are not suitable for calculation, so we must rewrite them. If we set  $r = \frac{p}{\sigma_x \sigma_y}$  we introduce the usual notation for the coefficient of correlation and write:

$$RR': \quad x = r \frac{\sigma_x}{\sigma_y} y$$

$$CC': \quad y = r \frac{\sigma_y}{\sigma_x} x$$

These are the same equations we arrived at in our general discussion above.

In this discussion we assumed that the means of the rows would fall on  $RR'$  and the means of the columns would fall on  $CC'$ . We must now consider the more usual situation where this does not occur.

If the values of  $x$  and  $y$  (the deviations from  $M_y$  and  $M_x$ ) be found for all associated pairs of values, then we find:



For the purpose of this paper, we shall assume that the function  $f(x, y)$  is continuous in the region  $R$  and that the partial derivatives  $f'_x$  and  $f'_y$  exist and are continuous in  $R$ . We shall also assume that the function  $f(x, y)$  is not constant in  $R$ .

Let us define the level curves of  $f(x, y)$  as the curves in  $R$  such that  $f(x, y) = c$ , where  $c$  is a constant.

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

$$\begin{aligned} \text{For all } (x, y) \text{ in } R, \quad f(x, y) &= c \\ \text{Differentiating both sides with respect to } x, \quad f'_x(x, y) &= 0 \\ \text{Differentiating both sides with respect to } y, \quad f'_y(x, y) &= 0 \end{aligned}$$

Therefore, the gradient vector  $\nabla f(P)$  is perpendicular to the level curve  $C$  at the point  $P$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .

Let  $C$  be a level curve of  $f(x, y)$  and let  $P$  be a point on  $C$ . Then the tangent line to  $C$  at  $P$  is perpendicular to the gradient vector  $\nabla f(P)$ .



$$A. (1) \leq (x - b, y)^2 = N \sigma_x^2 (1 - r^2)$$

$$\text{when } b, = r \frac{\sigma_x}{\sigma_y}$$

and where  $x$  is the actual deviation from the mean and  $b, y$  is the estimated deviation.

Proof.

$$\begin{aligned} (1) (x - b, y)^2 &= x^2 - 2xy + b,^2 y^2 \\ (2) \leq (x - b, y)^2 &\leq x^2 - 2b, xy + b,^2 y^2 \\ &= N \sigma_x^2 - 2b, N r + b,^2 N \sigma_y^2 \\ (3) &= N \sigma_x^2 - 2b, N b, \sigma_y^2 + b,^2 N \sigma_y^2 \\ (4) &= N \sigma_x^2 - N \sigma_y^2 b,^2 \\ (5) &= N \sigma_x^2 - N \sigma_y^2 \cdot r^2 \frac{\sigma_x^2}{\sigma_y^2} \\ (6) \leq (x - b, y)^2 &= N \sigma_x^2 (1 - r^2) \end{aligned}$$

Q.E.D.

Now if  $b,$  equals any other value such as  $b, = (r + \delta) \frac{\sigma_x}{\sigma_y}$ , then

$$B. \leq (x - b, y)^2 = N \sigma_x^2 (1 - r^2 + \delta^2)$$

Proof.

$$\begin{aligned} (1) \leq (x - b, y)^2 &= \leq \left[ x - (r + \delta) \frac{\sigma_x}{\sigma_y} y \right]^2 \\ (2) &= \leq \left[ x^2 - 2xy(r + \delta) \frac{\sigma_x}{\sigma_y} + (r + \delta)^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right] \\ (3) &= \leq \left[ x^2 - 2xyr \frac{\sigma_x}{\sigma_y} + r^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right] + \leq \left[ -2xy\delta \frac{\sigma_x}{\sigma_y} + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} y^2 + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \right] \\ (4) &= \leq \left( x - r \frac{\sigma_x}{\sigma_y} y \right)^2 - 2\delta \frac{\sigma_x}{\sigma_y} xy + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} y^2 + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} y^2 \\ (5) &= \leq \left( x - r \frac{\sigma_x}{\sigma_y} y \right)^2 - 2\delta \frac{\sigma_x}{\sigma_y} N r \sigma_x \sigma_y + 2r\delta \frac{\sigma_x^2}{\sigma_y^2} N \sigma_y^2 + \delta^2 \frac{\sigma_x^2}{\sigma_y^2} N \sigma_y^2 \\ (6) &= \leq \left( x - r \frac{\sigma_x}{\sigma_y} y \right)^2 - 2N r \delta \sigma_x^2 + 2N r \delta \sigma_x^2 + N \delta^2 \sigma_x^2 \\ (7) &= \leq \left( x - r \frac{\sigma_x}{\sigma_y} y \right)^2 + N \delta^2 \sigma_x^2 \\ (8) \leq (x - b, y)^2 &= N \sigma_x^2 (1 - r^2) + N \delta^2 \sigma_x^2 = N \sigma_x^2 (1 - r^2 + \delta^2) \end{aligned}$$

Now we showed when  $b, = r \frac{\sigma_x}{\sigma_y}$

$$A. \leq (x - b, y)^2 = N \sigma_x^2 (1 - r^2)$$

and when  $b, = (r + \delta) \frac{\sigma_x}{\sigma_y}$

$$B. \leq (x - b, y)^2 = N \sigma_x^2 (1 - r^2 + \delta^2)$$

The right-hand side of B is obviously greater than the right-hand side of A;

and where  $\epsilon$  is the actual deviation from the mean and  $\mu$  is the estimated deviation.

Proof.

- (1)  $\mu = \frac{1}{n} \sum_{i=1}^n x_i$
- (2)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$
- (3)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$
- (4)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$
- (5)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$
- (6)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$
- (7)  $\mu = \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + \mu$

So if  $\mu$  equals any other value such as  $\mu' \neq \mu$ , then

Proof.

- (1)  $\mu' = \frac{1}{n} \sum_{i=1}^n x_i$
- (2)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (3)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (4)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (5)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (6)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (7)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$
- (8)  $\mu' = \frac{1}{n} \sum_{i=1}^n (x_i - \mu') + \mu'$

Now we showed that

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

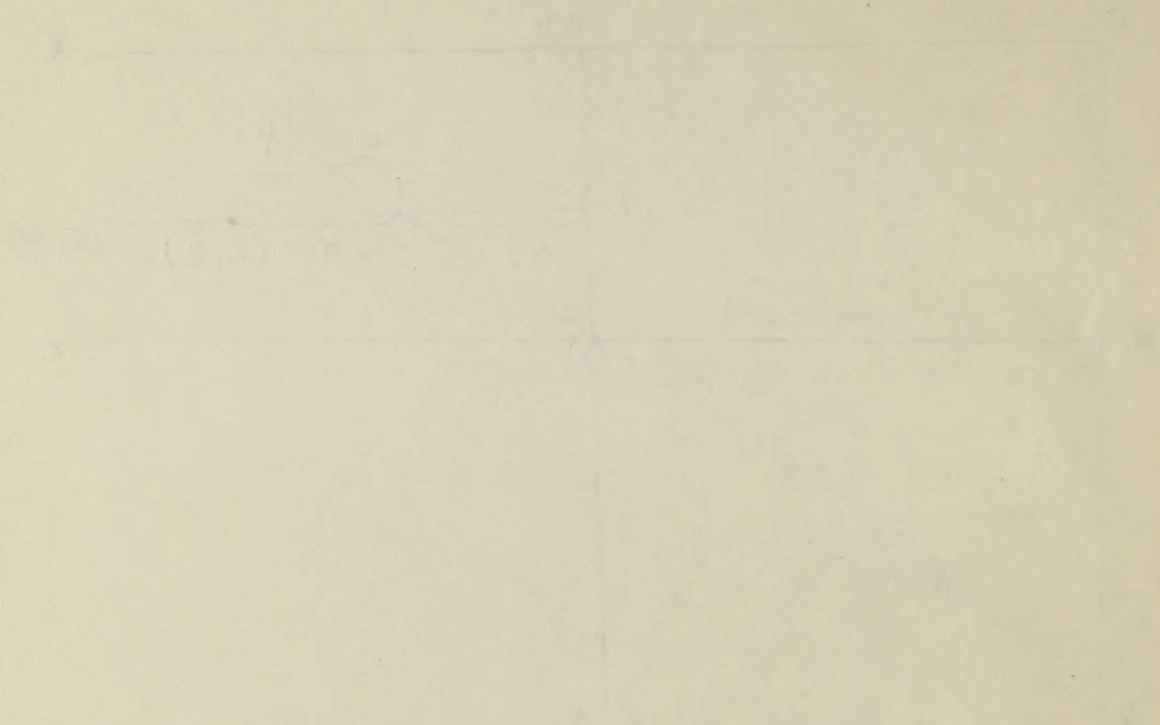
and when

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

The right-hand side of  $\mu$  is obviously greater than the right-hand side of  $\mu'$





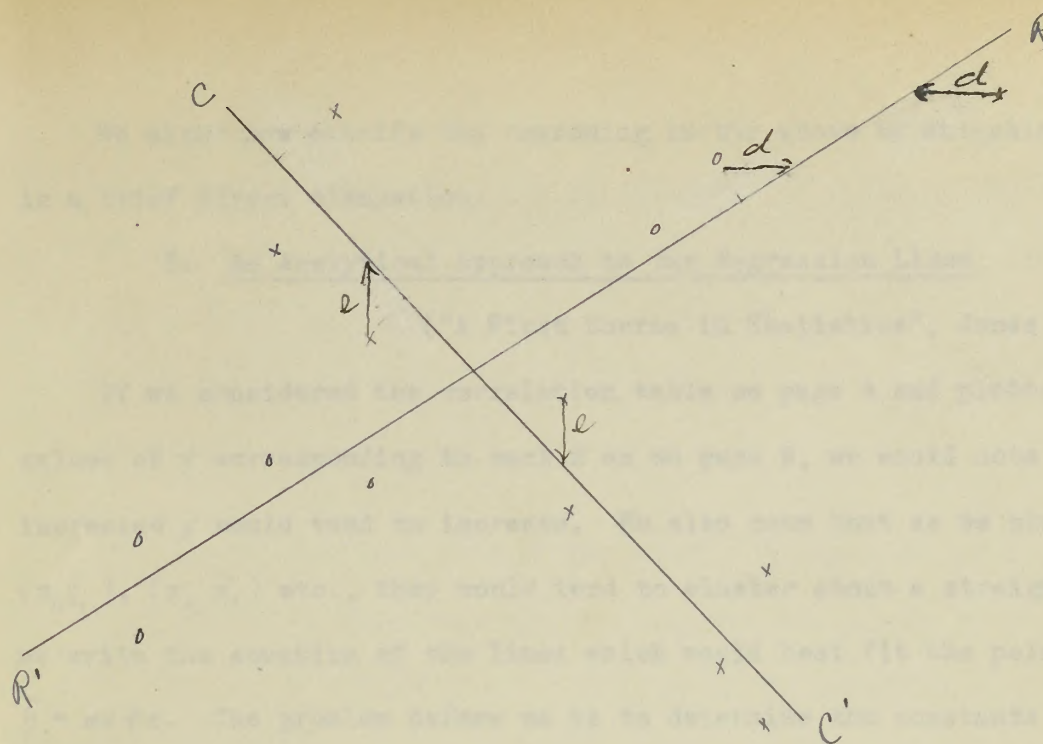


Let us consider the distribution in a row of type  $y$  with the origin at  $x$ .  
and find the root-mean-square deviation of the row about point  $x$  (or  $y$ ).  
See Note 1, page 1 of Introduction.

The root-mean-square deviation of the row about  $x$  is  
where  $n$  is the frequency of the row,  
is for the row  $x$  (or  $y$ ),  
where  $s_x$  is the standard deviation of the row,  
is for the row  $x$  (or  $y$ ),  
for the entire distribution then  
the  $s_y$  is the standard deviation of the row.

$s^2$  is the sum of the standard deviations of the rows and rows  
unweighted regardless of the slope of  $s^2$ , as a  $(x-y)$  row the only terms  
affected by  $s^2$ .  
Now  $s^2(x-y)$  is a minimum, so  $s^2$  must be a minimum for the  
value of  $s^2 = 0$ .





This says that the sum of the squares of the distances of the means of the rows from  $RR'$  (each multiplied by the frequency of that row) is the lowest possible when  $b_1 = r \frac{\sigma_x}{\sigma_y}$ .

The same can be proved in like manner for  $CC'$ ; that is, the sum of the squares of the distances of the means of the columns from  $CC'$  is the lowest possible when  $b_2 = r \frac{\sigma_y}{\sigma_x}$  and the equation of  $CC'$  is  $y = b_2 x$ .

Therefore, the equations

$$x = r \frac{\sigma_x}{\sigma_y} \quad \text{and} \quad y = r \frac{\sigma_y}{\sigma_x} \quad \text{may be regarded as}$$

"(a) equations for estimating each individual  $x$  from its associated  $y$  (and  $y$  from its associated  $x$ ) in such a way as to make the sum of the squares of errors of estimate the least possible, or (b) equations for estimating the mean of the  $x$ 's associated with a given type of  $y$  (and the mean of the  $y$ 's associated with a given type of  $x$ ) in such a way as to make the sum of the squares of errors of estimate the least possible when every mean is counted once for each observation on which it is based." Yule p. 72-3

These lines are called the lines of "best fit" of the actual lines of the means.

This says that the sum of the squares of the means of  
 the two lines EC (each multiplied by the frequency of that row) is the  
 least possible when  $x = \bar{x}$  and  $y = \bar{y}$ .  
 The same can be proved in like manner for EC; that is, the sum of the  
 squares of the distances of the means of the columns from EC is the least  
 possible when  $x = \bar{x}$  and the position of EC is  $y = \bar{y}$ .  
 Therefore, the equations  

$$x = \bar{x} \quad \text{and} \quad y = \bar{y}$$
 may be formulated as  
 "the equations for estimating each individual  $x$  from the associated  $y$   
 (and  $y$  from the associated  $x$ ) is such a way as to make the sum of the squares  
 of distances the least possible, for (a) equations for estimating the mean of  
 the  $x$ 's associated with a given type of  $y$  (and the mean of the  $y$ 's associated  
 with a given type of  $x$ ) is such a way as to make the sum of the squares of  
 errors of estimate the least possible when each error is counted once for  
 each observation on which it is based."  
 These lines are called the lines of "best fit" of the actual lines of  
 the means.



We might now clarify the reasoning in the above by attacking the problem in a brief direct discussion.

B. An Analytical Approach to the Regression Lines

("A First Course in Statistics", Jones P. 104 ff.)

If we considered the correlation table on page 4 and plotted the mean values of  $y$  corresponding to each  $x$  as on page 9, we would note that as  $x$  increased  $y$  would tend to increase. We also note that as we plot the points  $(x_1, \bar{y}_1)$ ,  $(x_2, \bar{y}_2)$  etc., they would tend to cluster about a straight line. If we write the equation of the lines which would best fit the points, it is  $\bar{y} = mx + c$ . The problem before us is to determine the constants  $m$  and  $c$  so that we may write this equation. If we can do this, we will be able to find the best average value of  $y$  corresponding to any  $x$ .

Now  $\bar{y}_1, \bar{y}_2, \bar{y}_3$  etc., were the best values of  $y$  corresponding to  $x_1, x_2, x_3$  etc., so if we rewrite the equation  $y = mx + c$  we will be still estimating the best  $y$  corresponding to any given  $x$  and basing our work on all the observations since  $\bar{y}$  is the best value of  $y$  in that particular column.

If  $x = x_1$ ,

$$y = mx_1 + c$$

But for any value  $x_1$  of  $x$  there may be several values of  $y$  as seen in the correlation table on page 4; if  $y_1$  is one of these values, the difference between it and the value given by the equation is  $(mx_1 + c) - y_1$ .

This difference measures the distance measured parallel to the  $y$  axis between the observed point  $(x_1, y_1)$  and the line  $y = mx + c$ . We now wish to find the equation of a line such that the sum of these differences for all paired values of  $x$  and  $y$  will be a minimum. Since some of these differences are positive and some negative, we will search for the equation which will make the sum of their squares a minimum. The problem then, is to find  $c$  and  $m$  which will make

$$(mx_1 + c - y_1)^2 + (mx_2 + c - y_2)^2 + \dots + (mx_n + c - y_n)^2$$

a minimum.

is a direct linear relation. The straight line passing through the origin is a direct linear relation.

2. An analytical expression for the regression line

(A. B. Stephens, "Statistics", pages 104-105.)  
If we consider the correlation table on page 1 and plot the values of  $y$  corresponding to each  $x$  as of page 1, we would have a scatter of points. If we draw a line through the points, we would have a regression line. If we write the equation of the line which would best fit the points, it is  $y = ax + b$ . The problem before us is to determine the constants  $a$  and  $b$  so that we may write this equation. If we can do this, we will be able to find the best average value of  $y$  corresponding to any  $x$ .

For  $y_1, y_2, \dots, y_n$ , the best value of  $y$  corresponding to  $x_1, x_2, \dots, x_n$  is the value of  $y$  which will be still satisfying the best  $y$  corresponding to any given  $x$  and having the same as all the observations on  $y$  in the best value of  $y$  in that particular column.

If  $x = x_1$   
 $y = y_1, y_2, \dots, y_n$   
For any value  $x$  of  $x$  there may be several values of  $y$  as seen in the correlation table on page 1; if  $y$  is one of these values, the difference between it and the value given by the equation is  $(y_1 - y)$ . This difference measures the distance measured parallel to the  $y$  axis between the observed point  $(x_1, y_1)$  and the line  $y = ax + b$ . We now wish to find the equation of a line such that the sum of these differences for all paired values of  $x$  and  $y$  will be a minimum. Since some of these differences are positive and some negative, we will search for the equation which will make the sum of their squares a minimum. The problem then is to find a

not a line will be a minimum.  
The equation of the line which will be a minimum is  $y = ax + b$  where  $a$  and  $b$  are given by the following formulas:  
$$a = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$
  
$$b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$$



If we consider this an expression in  $c$ , differentiate, and set equal to zero, we will have the value of  $c$  which makes the expression a minimum.

$$(mx_1 + c - y_1) + (mx_2 + c - y_2) + \dots + (mx_n + c - y_n) = 0$$

$$m(x_1 + x_2 + \dots + x_n) + nc - (y_1 + y_2 + \dots + y_n) = 0$$

$$m(n\bar{x}) + nc - n\bar{y} = 0$$

$$m\bar{x} + c - \bar{y} = 0$$

This equation passes through the point  $(\bar{x}, \bar{y})$ , the mean of the entire distribution. This suggests that we transpose the origin to the point  $(\bar{x}, \bar{y})$  so that  $x - \bar{x} = d$  and  $y - \bar{y} = k$ , now in the equation  $m\bar{x} + c - \bar{y} = 0$ , the value of  $c = 0$ .

Now returning to  $(mx_1 + c - y_1)^2 + \dots + (mx_n + c - y_n)^2$  and differentiating with respect to  $m$ .

$$x_1(mx_1 + c - y_1) + x_2(mx_2 + c - y_2) + \dots + x_n(mx_n + c - y_n) = 0$$

$$m(x_1^2 + x_2^2 + \dots + x_n^2) + c(x_1 + x_2 + \dots + x_n) - (x_1y_1 + x_2y_2 + \dots + x_ny_n) = 0$$

Now replacing the  $x$ 's and  $y$ 's by their deviations from the mean (i.e. transposing the origin to  $(\bar{x}, \bar{y})$ ).

$$m(d_1^2 + d_2^2 + \dots + d_n^2) + c(d_1 + d_2 + \dots + d_n) - (d_1k_1 + d_2k_2 + \dots + d_nk_n) = 0$$

$$m = \frac{d_1k_1 + d_2k_2 + \dots + d_nk_n}{d_1^2 + d_2^2 + \dots + d_n^2}$$

Now if  $P = \frac{d_1k_1 + d_2k_2 + \dots + d_nk_n}{n}$

and  $\sigma_x^2 = \frac{d_1^2 + d_2^2 + \dots + d_n^2}{n}$  as usual,

then  $m = \frac{nP}{n\sigma_x^2} = \frac{P}{\sigma_x^2}$

Thus we have shown that if we considered the equation  $y = mx + c$  and transferred the origin to  $\bar{x}, \bar{y}$  this equation would be

$$k = md + c$$

or

$$y - \bar{y} = m(x - \bar{x}) + c$$

If we consider the expression in (1), differentiating, and set equal to

zero, we will have the value of  $x$  which makes the expression a minimum.

$$f'(x) = 2x - 2 = 0 \quad (1)$$

$$2x - 2 = 0 \quad (2)$$

$$2x = 2 \quad (3)$$

$$x = 1 \quad (4)$$

This equation passes through the point  $(1, 1)$  the mean of the entire

distribution. This suggests that we transpose the origin to the point

$x = 1$  and  $y = 1$  and let  $x' = x - 1$  and  $y' = y - 1$ , then in the equation  $x' + y' = 0$ ,

the value of  $x' = 0$ .

Now returning to (1),  $f(x) = (x - 1)^2 + (y - 1)^2$  and differentiating

the right member to  $x'$ ,

$$f'(x') = 2x' = 0 \quad (5)$$

$$2x' = 0 \quad (6)$$

and replacing the  $x'$ 's and  $y'$ 's by their definitions from the mean (i.e., trans-

posing the origin to  $(1, 1)$ ,

$$f'(x) = 2(x - 1) = 0 \quad (7)$$

$$2(x - 1) = 0 \quad (8)$$

$$2x - 2 = 0 \quad (9)$$

$$2x = 2 \quad (10)$$

$$x = 1$$

$$x = 1 \quad (11)$$

then

$$x = 1 \quad (12)$$

Thus we have shown that if we considered the equation  $y = x + c$  and

transferred the origin to  $(1, 1)$ , this equation would be

$$y' = x' + c \quad (13)$$

or

$$y' = x' + c \quad (14)$$



and  $C = 0$ ,  $m = \frac{P}{\sigma_x^2}$

$$\therefore (y - \bar{y}) = \frac{P}{\sigma_x^2} (x - \bar{x})$$

If  $x - \bar{x} = 1$ , then  $y - \bar{y} = \frac{P}{\sigma_x^2}$ , so  $\frac{P}{\sigma_x^2}$  measures the change in the deviation of  $y$  corresponding to a unit change in the deviation of  $x$ .

If we repeated the entire discussion interchanging the  $x$ 's and  $y$ 's we would arrive in exactly the same steps at the result

$$(x - \bar{x}) = \frac{P}{\sigma_y^2} (y - \bar{y})$$

Thus if  $(y - \bar{y}) = 1$ ,  $(x - \bar{x}) = \frac{P}{\sigma_y^2}$ , so  $\frac{P}{\sigma_y^2}$  measures the deviation in  $x$  from the mean of  $x$  corresponding to a unit deviation in  $y$  from the mean of  $y$ .

Therefore, either  $\frac{P}{\sigma_x^2}$  or  $\frac{P}{\sigma_y^2}$  may be considered as good measures for the correlation between  $x$  and  $y$ ; they are not alike because  $\frac{P}{\sigma_x^2}$  gives the change in  $y$  corresponding to a unit change in  $x$  and  $\frac{P}{\sigma_y^2}$  gives the change in  $x$  corresponding to a unit change in  $y$ . If we wish to compare these changes, we must reduce them to ratios which will be comparable, so we divide  $x - \bar{x}$  by the standard deviation  $\sigma_x$  and  $y - \bar{y}$  by  $\sigma_y$  and compare

$$y - \bar{y} = \frac{P}{\sigma_x^2} (x - \bar{x})$$

Divide both sides by  $\sigma_y$

$$\frac{y - \bar{y}}{\sigma_y} = \frac{P}{\sigma_x \sigma_y} \left( \frac{x - \bar{x}}{\sigma_x} \right)$$

Similarly for

$$x - \bar{x} = \frac{P}{\sigma_y^2} (y - \bar{y})$$

Divide by  $\sigma_x$

$$\frac{x - \bar{x}}{\sigma_x} = \frac{P}{\sigma_x \sigma_y} \left( \frac{y - \bar{y}}{\sigma_y} \right)$$

Now we have  $\frac{P}{\sigma_x \sigma_y}$  as the measure of correlation and write  $r = \frac{P}{\sigma_x \sigma_y}$

Now substituting  $r$  in our equations, we have

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

which are the equations of the lines of regression of  $y$  on  $x$  and  $x$  on  $y$

and

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

If  $x = 1$ , then  $y = 1$ , so  $\bar{x} = 1$ ,  $\bar{y} = 1$ , and the change in

the deviation of  $y$  corresponding to a unit change in the deviation of  $x$ .

If we consider the entire distribution, the  $x$ 's and  $y$ 's are

would give us exactly the same result.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Thus if  $y = 1$ ,  $\bar{y} = 1$ ,  $(x - \bar{x}) = 0$ , so  $\bar{x} = 1$ , and the deviation

in  $x$  from the mean of  $x$  corresponding to a unit deviation in  $y$  from the mean

of  $y$ .

Therefore, either  $r$  or  $\frac{1}{r}$  may be considered as good measures for

the correlation between  $x$  and  $y$ ; they are not alike because  $\frac{1}{r}$  gives the

change in  $y$  corresponding to a unit change in  $x$  and  $\frac{1}{r}$  gives the change in

$x$  corresponding to a unit change in  $y$ . If we wish to compare these values,

we must reduce them to ratios which will be comparable, so we divide  $x - \bar{x}$

by the standard deviation of  $x$  and  $y - \bar{y}$  by the standard deviation of  $y$ .

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}}{\sqrt{\sum \left( \frac{(x - \bar{x})}{\sigma_x} \right)^2 \sum \left( \frac{(y - \bar{y})}{\sigma_y} \right)^2}}$$

Divide both sides by

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}}{\sqrt{\sum \left( \frac{(x - \bar{x})}{\sigma_x} \right)^2 \sum \left( \frac{(y - \bar{y})}{\sigma_y} \right)^2}}$$

Divide both sides by

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}}{\sqrt{\sum \left( \frac{(x - \bar{x})}{\sigma_x} \right)^2 \sum \left( \frac{(y - \bar{y})}{\sigma_y} \right)^2}}$$

Divide both sides by

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}}{\sqrt{\sum \left( \frac{(x - \bar{x})}{\sigma_x} \right)^2 \sum \left( \frac{(y - \bar{y})}{\sigma_y} \right)^2}}$$

Now we have the measure of correlation and with

the substituting  $r$  in our equation, we have

$$r = \frac{\sum \frac{(x - \bar{x})}{\sigma_x} \frac{(y - \bar{y})}{\sigma_y}}{\sqrt{\sum \left( \frac{(x - \bar{x})}{\sigma_x} \right)^2 \sum \left( \frac{(y - \bar{y})}{\sigma_y} \right)^2}}$$

which are the equations of the lines of regression of  $y$  on  $x$  and  $x$  on  $y$ .



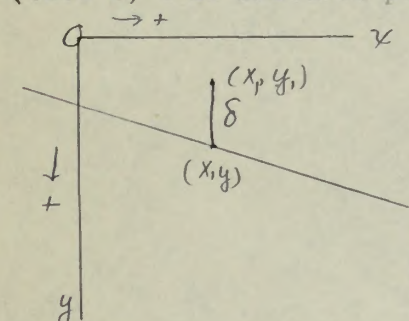
respectively.

C. A Development of the Regression Lines Introduced to Show the Range of Values of  $r$ . An Exact Definition of the Correlation Coefficient.

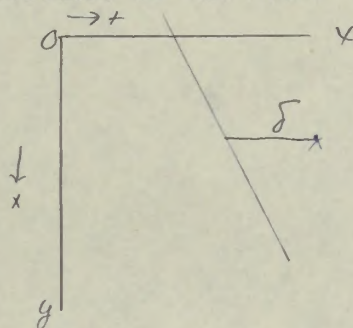
("Mathematical Part of Elementary Statistics" Camp)

If we think of the general case of correlation, we could think of the data represented by dots spread over the paper. We wish to find the equation of that straight line which, on the whole, will come nearest to all these dots. That is, if we let  $\delta$  be the distance between a dot and the line, we wish to make  $\sum \delta^2$  a minimum.

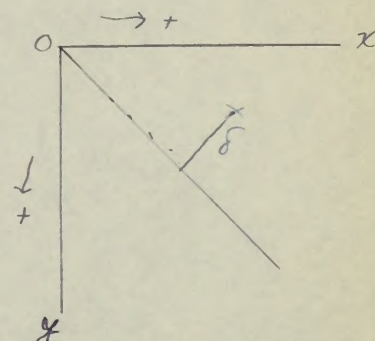
There are three cases, depending on whether (case a)  $d$  is measured parallel to the  $y$  axis; (case b)  $d$  is measured parallel to the  $x$  axis; (case c)  $d$  is measured perpendicular to the line.



(a) The regression line of  $y$  on  $x$ .



(b) The regression line of  $x$  on  $y$ .



(c) Camp calls this the "geometrically best fitting line."

A Method of finding Lines (a) and (c) by Means of Least Squares - Introduced to Show the Range of  $r$ .

Case (a) To obtain the equation of the regression of  $\bar{y}$  on  $\bar{x}$ .

Here we wish  $\sum \delta^2 f$  to be a minimum, where  $f$  represents the frequency.

$$(1) \delta = y - y_p,$$

$$(2) \text{ Let } y = A + Bx,$$

$$(3) \text{ Then } \delta = A + Bx - y,$$

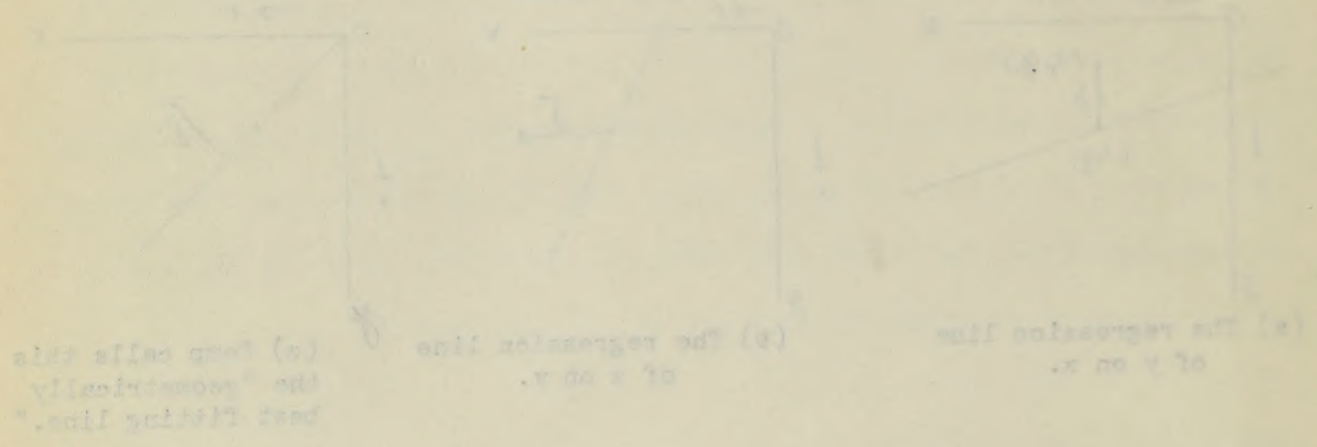
$$(4) \delta^2 = A^2 + B^2 x^2 + y^2 + 2ABx - 2Ay - 2Bxy,$$

$$(5) \frac{1}{N} \sum \delta^2 f = \frac{1}{N} \sum (A^2 + B^2 x^2 + y^2 + 2ABx - 2Ay - 2Bxy) f(x, y)$$

2. A Derivation of the Regression Lines Introduced to Show the  
Meaning of Values of  $r$ . The Exact Definition of the Correlation  
Coefficient.  
[Statistical Part of Elementary Statistics, Chap.]

If we think of the general case of correlation, we could think of the data represented by dots spread over the paper. We wish to find the equation of that straight line which, on the whole, will come nearest to all these dots. That is, if we let  $d$  be the distance between a dot and the line, we wish to make  $\sum d^2$  a minimum.

There are three cases, depending on whether (case a)  $d$  is measured parallel to the  $y$  axis; (case b)  $d$  is measured parallel to the  $x$  axis; (case c)  $d$  is measured perpendicular to the line.



A Method of Finding Lines (a) and (c) by Means of Least Squares - Introduced to Show the Meaning of  $r$ .

Case (a) To obtain the equation of the regression of  $y$  on  $x$ .

Let us write  $\sum d^2$  to be a minimum, where  $d$  represents the frequency.

$$\begin{aligned} (1) \quad \sum d^2 &= \sum (y - \bar{y})^2 \\ (2) \quad \text{Let } y &= A + Bx \\ (3) \quad \text{Then } \sum d^2 &= \sum (y - A - Bx)^2 \\ (4) \quad \sum d^2 &= \sum (y^2 - 2Ay - 2Bxy + A^2 + 2ABx + B^2x^2) \\ (5) \quad \frac{1}{2} \frac{d}{dA} \sum d^2 &= \sum (-y + A + Bx) = 0 \end{aligned}$$



$$(6) \frac{1}{N} \sum \delta^2 f = \frac{A^2}{N} \sum f(x, y) + \frac{B^2}{N} \sum x^2 f(x, y) + \frac{1}{N} \sum y^2 f(x, y) + 2 \frac{AB}{N} \sum x f(x, y) - 2 \frac{A}{N} \sum y f(x, y) - 2 \frac{B}{N} \sum x y f(x, y)$$

$$(7) \frac{1}{N} \sum \delta^2 f = A^2 + B^2 \sigma_x^2 + \sigma_y^2 - 2 B r \sigma_x \sigma_y$$

(8) The expression on the left of (7) is the sum of squares and is positive, so the expression on the right is positive. So if we take  $A = 0$ , we can then find what value of  $B$  will make this expression a minimum.

That is, we wish to make

$$(9) \sigma_y^2 + \sigma_x^2 (B^2 - 2 B r \frac{\sigma_y}{\sigma_x}) \text{ a minimum.}$$

(10) This expression will be a minimum when  $B^2 - 2 B r \frac{\sigma_y}{\sigma_x}$  is a minimum.

(11) Differentiating with respect to  $B$  and equaling to zero, we get

$$2 B - 2 r \frac{\sigma_y}{\sigma_x} = 0$$

$$(12) B = \frac{r \sigma_y}{\sigma_x}$$

(13) Substituting (12) in (2) and noting that  $A = 0$  we get

$$y = \frac{r \sigma_y}{\sigma_x} x$$

This is the equation for the regression of  $\bar{y}$  on  $\bar{x}$

#### Case (b)

In the same manner interchanging the  $y$ 's and  $x$ 's we get the equation of the regression of  $\bar{x}$  on  $\bar{y}$

$$x = \frac{r \sigma_x}{\sigma_y} y$$

(14) Now in (7) we had

$$\frac{1}{N} \sum \delta^2 f = \sigma_y^2 + A^2 + B^2 \sigma_x^2 - 2 B r \sigma_y \sigma_x$$

$$\text{but } A = 0 \text{ and } B = r \frac{\sigma_y}{\sigma_x}$$

$$(15) \text{ Hence } \frac{1}{N} \sum \delta^2 f = \sigma_y^2 + r^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 - 2 r^2 \frac{\sigma_y}{\sigma_x^2} \sigma_y \sigma_x^2$$

$$(16) \frac{1}{N} \sum \delta^2 f = \sigma_y^2 (1 - r^2)$$

Now the left side of this equation is positive, since it is the sum of squares, so the right side is positive. Hence

$$1 - r^2 \geq 0$$

and

$$-1 \leq r \leq 1$$





Also in similar manner in the case of the equation for the regression of  $\bar{Y}$  on  $\bar{X}$  we may show

$$\frac{1}{N} \sum \delta^2 f = \sigma_x^2 (1 - r^2)$$

$$\text{and } -1 \leq r \leq 1$$

Now to return to Case (c) and to find the equation of the "Geometrically best-fitting line."

- (1) In Analytic Geometry, the formula for the distance from the point

$(x_1, y_1)$  to the line  $Ax + By + C = 0$  is

$$d = \frac{Ax_1 + By_1 + C}{\pm \sqrt{A^2 + B^2}} \quad \text{where } d \text{ is positive.}$$

- (2) Our equation is

$$y = A + Bx \quad \text{or} \quad Bx - y + A = 0 \quad \text{and}$$

the point is  $(x_1, y_1)$

(3) Hence 
$$\delta = \frac{Bx_1 - y_1 + A}{\pm \sqrt{B^2 + 1}}$$

(4) 
$$\frac{1}{N} \sum \delta^2 f = \frac{1}{N} \sum \left( \frac{Bx - y + A}{\pm \sqrt{B^2 + 1}} \right)^2 f(x, y)$$

(5) 
$$\frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} [B^2 \sigma_x^2 + \sigma_y^2 + A^2 - 2B \bar{P}_{xy}]$$

(6) 
$$\frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} [B^2 \sigma_x^2 + \sigma_y^2 + A^2 - 2Br \sigma_x \sigma_y]$$

- (7) Here again, the left side is positive, so the right side is positive.

If we let  $A = 0$ , we may solve for that value of  $B$  which will make

$$\frac{B^2 \sigma_x^2 + \sigma_y^2 - 2Br \sigma_x \sigma_y}{B^2 + 1} \quad \text{a minimum.}$$

- (8) In this problem, however, we are interested only when the standard

deviation is used as a unit, so we first set  $\sigma_x = \sigma_y = 1$

- (9) Therefore (7) becomes

$$\frac{B^2 + 1 - 2Br}{1 + B^2}$$

and we wish to find the value of  $B$  which will make this a minimum.

- (10) Rewrite (9)

$$1 - \frac{2Br}{B^2 + 1}$$

Also in similar manner in the case of the equation for the trajectory of  
 we have seen that  
 and  
 Now to return to Case (a) and to find the equation of the trajectory  
 postulating that

(1) In the velocity component, the formula for the distance from the point  
 to the line  $W$  is  $\frac{1}{2} W^2$  is  
 where  $W$  is constant.  
 (2) The equation is  
 and  
 The solution is  $Y$

(3) Hence  $\frac{dY}{dX} = \frac{W^2 - Y^2}{2Y}$   
 (4)  $\frac{dY}{dX} = \frac{W^2 - Y^2}{2Y}$   
 (5)  $\frac{dY}{dX} = \frac{W^2 - Y^2}{2Y}$

(6) Now again, the left side is constant, so the right side is constant.  
 If we let  $A = 0$ , we may solve for that value of  $W$  which will make  
 a minimum.  
 (7) In this problem, however, we are interested only when the standard  
 deviation is used as a unit, so we first set  
 (8) Therefore  $Y$  becomes

and we wish to find the value of  $W$  which will make this a minimum.  
 (9) Therefore (a)  
 (10)  $\frac{dY}{dX} = \frac{W^2 - Y^2}{2Y}$



(11) Differentiating with respect to  $B$  and equating to zero

$$B = \pm 1$$

(12) When  $r > 0$

$$1 - \frac{2Br}{1+B^2} \text{ will be a minimum when } B = +1$$

When  $r < 0$

$$1 - \frac{2Br}{1+B^2} \text{ will be a minimum when } B = -1$$

When  $r = 0$

$$1 - \frac{2Br}{1+B^2} = 1 \text{ and cannot be a minimum.}$$

(13) So we may write the equation for the "geometrically best-fitting line"

$$y = x \quad \text{if } r > 0 \quad \text{and } \sigma_x = \sigma_y = 1$$

$$y = -x \quad \text{if } r < 0 \quad \text{and } \sigma_x = \sigma_y = 1$$

Now if we let  $\sigma_x = \sigma_y = 1$  and rewrite the equations for the regression lines, we have

(a) The equation of the regression of  $\bar{y}$  on  $\bar{x}$

$$y = rx$$

(b) The equation of the regression of  $\bar{x}$  on  $\bar{y}$

$$x = ry$$

(c) The equation of the "geometrically best-fitting" line

$$y = x \quad \text{if } r > 0$$

$$y = -x \quad \text{if } r < 0$$

(d) When  $\delta$  is measured parallel to the  $y$  axis  $\frac{1}{N} \sum \delta^2 f = 1 - r^2$

(e) When  $\delta$  is measured parallel to the  $x$  axis  $\frac{1}{N} \sum \delta^2 f = 1 - r^2$

(11) The regression line with respect to  $Y$  and equation to zero

$$B = 0$$

(12) The regression line with respect to  $X$  and equation to zero

$$A = 0$$

and  $B > 0$

$$B = 0$$

and  $A < 0$

$$A = 0$$

and cannot be a minimum.

(13) To be written for the "geometrically best-fitting line"

$$Y = 0.5X + 1$$

$$Y = 0.5X + 1$$

For  $Y$  we have  $Y = 0.5X + 1$  and write the equations for the

regression lines, we have

(a) The equation of the regression of  $Y$  on  $X$

$$Y = 0.5X + 1$$

(b) The equation of the regression of  $X$  on  $Y$

$$X = 2Y - 2$$

(c) The equation of the "geometrically best-fitting" line

$$Y = 0.5X + 1$$

$$Y = 0.5X + 1$$

(d) The line is parallel to the  $Y$  axis  $X = 0$

(e) The line is parallel to the  $X$  axis  $Y = 0$



(f) When  $\delta$  is the perpendicular distance from the point to the line

$$\frac{1}{N} \sum \delta^2 f = \frac{1}{B^2 + 1} (1 + B^2 - 2Br)$$

$$\text{when } r > 0 \quad B = 1$$

$$r < 0 \quad B = -1$$

$$\text{so } \frac{1}{N} \sum \delta^2 f = \frac{2 - 2r}{2} = 1 - r \quad r > 0$$

$$\frac{1}{N} \sum \delta^2 f = \frac{2 + 2r}{2} = 1 + r \quad r < 0$$

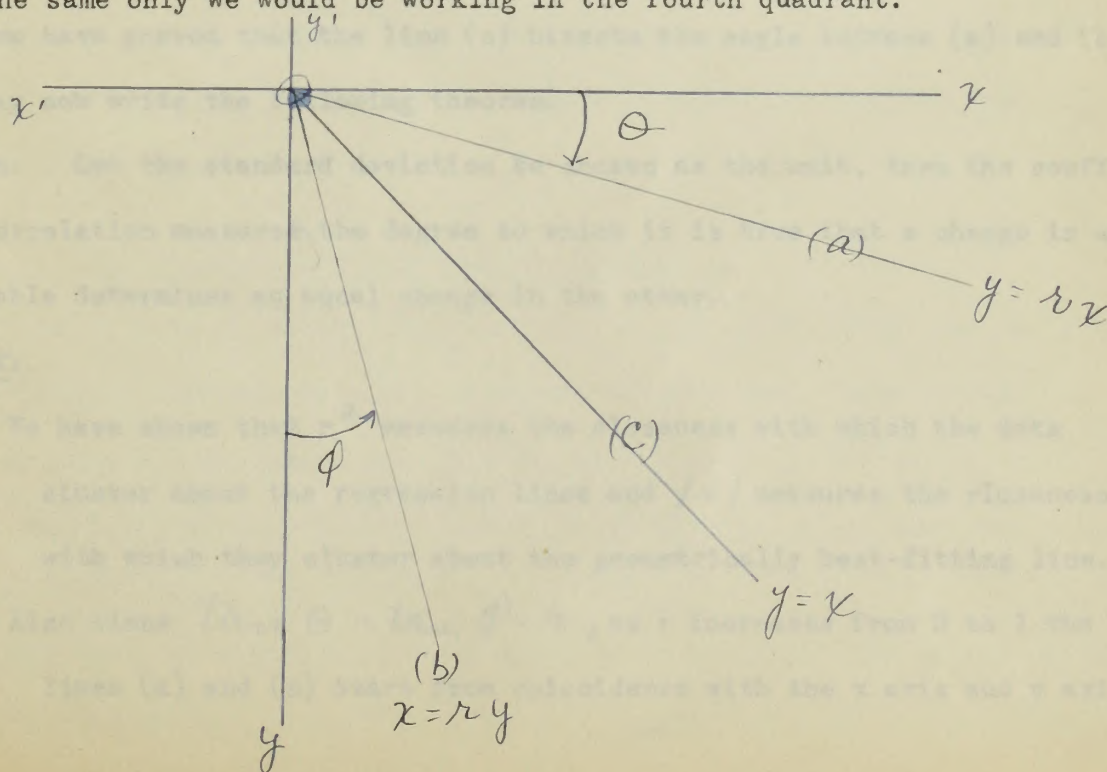
Therefore, from (d), (e), and (f) we may write

Th.  $|r|$  measures the closeness with which the dots cluster about the geometrically best-fitting line;  $r^2$  measures the closeness with which the dots cluster about the regression lines (distances in the last case being measured parallel to the y and x axis).

Now we are in a position to prove the following theorem.

Th. If  $\sigma_x = \sigma_y = 1$ , the line  $y = x$ , when  $r > 0$ , bisects the angle between the lines  $y = rx$  and  $x = ry$ ; when  $r < 0$ , the line  $y = -x$  bisects the angle between the lines  $y = rx$  and  $x = ry$ .

We shall prove the first part; for the proof of the second part would be the same only we would be working in the fourth quadrant.



(7) Let  $d$  be the perpendicular distance from the point to the line

$$Ax + By + C = 0$$

then

$$d = \frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}}$$

$$\text{so } d = \frac{|A(-1) + B(1) + C|}{\sqrt{A^2 + B^2}} = \frac{|-A + B + C|}{\sqrt{A^2 + B^2}}$$

$$= \frac{|-1 + 1 + 1|}{\sqrt{1^2 + 1^2}} = \frac{1}{\sqrt{2}}$$

Therefore, from (4), (5), and (7) we may write

7.  $d_1$  measures the distance from the point cluster

about the geometrically best-fitting line;  $d_2$  measures the distance from

which the data cluster about the regression line (distance in the first case

being measured parallel to the  $y$  and  $x$  axis).

Now we are in a position to prove the following theorem.

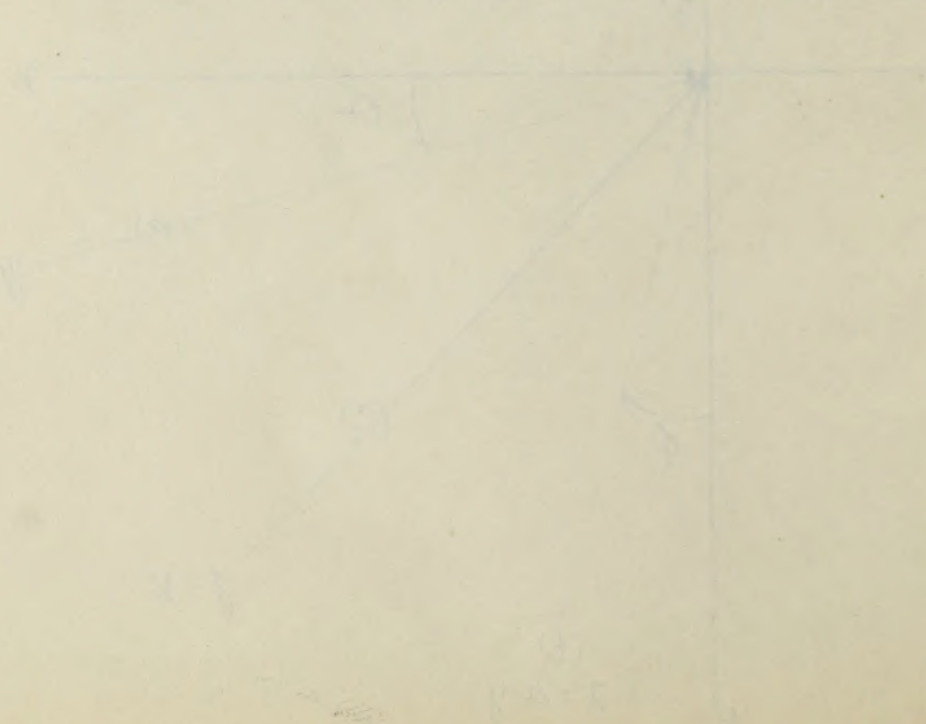
8. If  $d_1 = d_2$ , the line  $y = x$ , when  $d_1 = d_2$ , bisects the

angle between the lines  $y = x$  and  $y = -x$ ; and if  $d_1 < d_2$ , the line

$y = x$  bisects the angle between the lines  $y = x$  and  $y = -x$ .

We shall prove the first part, for the proof of the second part would

be the same only we would be working in the fourth quadrant.





Proof:

(1) Since there is no constant term, these lines all pass through the origin.

(Note the positive direction of  $y$  is downward)

(2) The equation for the line (a) is

$$y = r x$$

$$\text{so } r = \frac{y}{x}$$

$$\therefore r = \tan \theta$$

(3) Similarly for (b)  $r = \frac{x}{y}$

$$\text{so } r = \tan \phi$$

(4)  $\therefore$  from (2) and (3)  $\tan \phi = \tan \theta$

$$\therefore \phi = \theta$$

(5) The line  $y = x$  bisects the angle  $x$  o  $y$

(6) So the angle between (a) and (c) is

$$45^\circ - \theta$$

The angle between (b) and (c) is

$$45^\circ - \phi$$

(7)  $\therefore$  Since  $\phi = \theta$

$$45^\circ - \phi = 45^\circ - \theta$$

And we have proved that the line (c) bisects the angle between (a) and (b).

We may now write the following theorem.

Th. Let the standard deviation be chosen as the unit, then the coefficient of correlation measures the degree to which it is true that a change in one variable determines an equal change in the other.

Proof:

(1) We have shown that  $r^2$  measures the closeness with which the dots

cluster about the regression lines and  $|r|$  measures the closeness

with which they cluster about the geometrically best-fitting line.

(2) Also since  $\tan \theta = \tan \phi = r$ , as  $r$  increases from 0 to 1 the

lines (a) and (b) start from coincidence with the  $x$  axis and  $y$  axis,

Proof:

(1) Since there is no constant term, these lines all pass through the origin.

(2) Since the positive direction of  $y$  is upwards

(3) The equation for the line (a) is

$$y = x$$

so

$$\tan \theta = \frac{y}{x} = 1$$

(4) Similarly for (b)  $y = 2x$

$$\tan \theta = 2$$

$$\tan \theta = 2$$

(5) From (3) and (4)  $\tan \theta = 1$  and  $\tan \theta = 2$

$$\theta = 45^\circ$$

(6) The line  $y = x$  bisects the angle  $x$  and  $y$

(7) So the angle between (a) and (c) is

$$45^\circ - 0^\circ = 45^\circ$$

The angle between (b) and (c) is

$$45^\circ - 45^\circ = 0^\circ$$

(8) Since  $\theta = 45^\circ$

$$\theta = 45^\circ$$

and we have proved that the line (c) bisects the angle between (a) and (b).

We may now write the following theorem.

Theorem. Let the standard deviation be measured the unit, then the coefficient

of correlation measures the degree to which it is true that a change in one

variable determines an equal change in the other.

Proof:

(1) We have shown that  $r$  measures the closeness with which we have

cluster about the regression lines and  $r$  measures the closeness

with which they cluster about the geometrically best-fitting line.

(2) Also since  $r = \cos \theta$ , as  $r$  increases from 0 to 1 the

lines (a) and (b) start from coincidences with the  $x$  and  $y$  axes.



respectively, and rotate with equal angular velocities in the direction of  $c$ . When  $r = 1$ , they coincide with  $c$ .

(Note when  $r = -1$ , these lines coincide with  $(y = -x)$ )

- (3) For points on  $c$  a change in one variable determines an equal change in the other.

For the slope of the line is 1 and if  $x'', y''$  and  $x', y'$  are two points on it

$$\frac{y'' - y'}{x'' - x'} = 1$$

$$y'' - y' = x'' - x'$$

- (4) Therefore, the larger  $r$  is, the nearer (a) and (b) come to coincidence with (c); the nearer the dots lie to  $c$ ; and the nearer we have the condition that an equal change in one variable produces an equal change in the other.

We might now sum this discussion with the following statement of the above theorem.

Th. The coefficient of correlation measures the degree to which it is true that a relative change in one variable determines an equal relative change in the other. By a relative change is meant the ratio of the absolute change to the standard deviation.

#### D. The Standard Deviation of the Arrays

When we have found the equations of the regression lines, we are interested in knowing the dispersion of the rows and columns about these lines.

On page 11, we found that  $\sum (x - b, y)^2 = N \sigma_x^2 (1 - r^2)$  so we might define the standard deviation of the rows about the regression line of  $x$  on  $y$  as  $\sigma_x \sqrt{1 - r^2}$

Similarly we might define the standard deviation of the columns about the regression line of  $y$  on  $x$  as  $\sigma_y \sqrt{1 - r^2}$

respectively, and rotate with equal angular velocities in the direction

of  $\omega$ . When  $\omega = 1$ , they coincide with  $\omega$ .

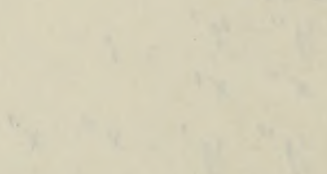
(Note when  $\omega = -1$ , these lines coincide with  $\omega = -1$ .)

(2) For points on a straight line in one variable determine an equal change in

the other.

For the slope of the line is  $\frac{1}{\omega}$  and if  $\omega = 1$  and  $\omega = -1$  are two

points on it



(3) Therefore, the points (1, 1) and (2, -1) are in collineation

with (1, 1) and (2, -1) and the points (1, 1) and (2, -1) are

collinear and an equal change in one variable produces an equal

change in the other.

We might now sum this discussion with the following statement of the

above theorem.

11. The coefficient of correlation measures the degree to which it is

true that a relative change in one variable determines an equal relative

change in the other. If a relative change is made the ratio of the absolute

change to the standard deviation.

### 12. The Standard Deviation of the Array

When we have found the equation of the regression line, we are interested

in finding the standard deviation of the rows and columns about these lines.

As was shown, we found that  $\hat{y} = a + bx$  and  $\hat{x} = c + dy$ .

So we might define the standard deviation of the rows about the regression

line of  $y$  on  $x$  as  $\sigma_y$ .

Similarly we might define the standard deviation of the columns about the

regression line of  $x$  on  $y$  as  $\sigma_x$ .



## Chapter IV

The Correlation Surface

(Elderton - "Frequency, Curves and Correlation")  
 (Forsyth - "Mathematical Analysis of Statistics")

The equation representing the correlation surface where the probability of the joint occurrence of deviations of  $x$  and  $y$  from their respective means, whether they are dependent or independent, can be written

$$Z = K e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

This equation is developed by Elderton in his "Frequency, Curves and Correlation" and is assumed in this discussion.

Our interest now is to replace the constants in the equation by expressions which will be of service in interpreting correlation tables; i.e. the standard deviations of  $x$  and  $y$  and the coefficient of correlation. This can be done by finding the volume  $N$  under the surface  $Z$ .

$$\therefore N = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Z dx dy$$

Let us consider first  $\int_{-\infty}^{\infty} Z dx$

$$\begin{aligned} (1) &= \int_{-\infty}^{\infty} K e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx \\ (2) &= K \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + \frac{h^2y^2}{a} - \frac{h^2y^2}{a} + by^2)} dx \\ (3) &= K e^{-\frac{a}{2}(\frac{by^2}{a} - \frac{h^2y^2}{a^2})} \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{hy}{a})^2} dx \end{aligned}$$

$$(4) \text{ Now evaluate } \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{hy}{a})^2} dx$$

$$\text{Let } X = x + \frac{hy}{a} \quad dX = dx$$

$$\int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{hy}{a})^2} dx = \frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{a}{2}X^2} dX$$

Since we will show later that any cross section of the normal surface made by a plane parallel to the  $yZ$  plane or parallel to the  $xZ$  plane is a normal curve and is symmetrical,  $2 \int_0^{\infty} e^{-\frac{a}{2}X^2} dX = \int_{-\infty}^{\infty} e^{-\frac{a}{2}X^2} dX$

$$(5) \therefore \int_{-\infty}^{\infty} e^{-\frac{a}{2}(x + \frac{hy}{a})^2} dx = \frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{a}{2}X^2} dX$$

(Latterly - "Correlation Surfaces and Correlation")  
 (Formerly - "Statistical Analysis of the Correlation")

The question now arises: how can the correlation surface be determined? The answer is: by the method of least squares. This method is based on the assumption that the correlation surface is a smooth function of the variables  $x$  and  $y$ . The method of least squares is a method of fitting a surface to a set of data points. It is a method of minimizing the sum of the squares of the residuals. The residuals are the differences between the observed values and the values predicted by the surface. The method of least squares is a method of finding the surface that best fits the data points.

This question is answered by Kistner in his "Theory of Correlation Surfaces and Correlation". It is assumed in this discussion.

Our interest now is to obtain the equation of the correlation surface. This can be done by fitting the surface to the data points. The method of least squares is a method of finding the surface that best fits the data points. It is a method of minimizing the sum of the squares of the residuals. The residuals are the differences between the observed values and the values predicted by the surface. The method of least squares is a method of finding the surface that best fits the data points.

$$\begin{aligned} (1) \quad & \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - a - bx - cy)^2 = 0 \\ (2) \quad & \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - a - bx - cy)^2 = 0 \\ (3) \quad & \frac{\partial}{\partial c} \sum_{i=1}^n (y_i - a - bx - cy)^2 = 0 \end{aligned}$$

Since we will show later that any curve section of the correlation surface is a curve of the type  $y = a + bx + cy$ , the equation of the correlation surface can be written in the form

$$y = a + bx + cy$$



$$(6) \text{ So } \int_{-\infty}^{\infty} z \, dx = k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2} \left(b - \frac{k^2}{a}\right)}$$

Since we integrated  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \, dx \, dy$  with respect to  $x$  first, we have then in  $k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2} \left(b - \frac{k^2}{a}\right)}$  a typical  $y$  section; so we may write

$$j = k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2} \left(b - \frac{k^2}{a}\right)}$$

where  $j$  is the frequency of this particular section.

Since it may be shown that a normal curve can be written  $j = m e^{-\frac{y^2}{2} \cdot \frac{1}{\sigma_y^2}}$

we have

$$(7) \quad \frac{1}{\sigma_y^2} = b \left(1 - \frac{k^2}{ab}\right)$$

(8) Now if we integrate  $\int_{-\infty}^{\infty} z \, dy$  in exactly the same manner we find

$$\frac{1}{\sigma_x^2} = a \left(1 - \frac{k^2}{ab}\right)$$

$$(9) \text{ Let } r = -\frac{k}{\sqrt{ab}}, \quad r^2 = \frac{k^2}{ab}$$

$$(10) \text{ From (7) } \frac{1}{\sigma_y^2} = b(1-r^2), \quad b = \frac{1}{\sigma_y^2(1-r^2)}$$

$$(11) \text{ From (8) } \frac{1}{\sigma_x^2} = a(1-r^2), \quad a = \frac{1}{\sigma_x^2(1-r^2)}$$

$$(12) \text{ In (9) } k = -r \sqrt{ab}$$

$$\text{so } k = \frac{-r}{\sigma_x \sigma_y (1-r^2)}$$

We now have the constants  $a$ ,  $b$ , and  $h$  expressed in the desired terms; only  $k$  remains.

Returning to

$$N = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \, dx \, dy$$

and substituting for  $\int_{-\infty}^{\infty} z \, dx$  its value  $k \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2\sigma_y^2}}$ , we have

$$(13) \quad N = k \sqrt{\frac{2\pi}{a}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma_y^2}} dy$$

$$(14) \quad \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma_y^2}} dy = \sigma_y \sqrt{2\pi}$$

$$(15) \quad \therefore N = k \sqrt{\frac{2\pi}{a}} \cdot \sigma_y \sqrt{2\pi} = \frac{2\pi k \sigma_y}{\sqrt{a}}$$

$$(1) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

Since we are interested in the first two terms of the expansion, we may write

in the frequency of this particular section.

There is no need to show that a normal wave can be written

as

$$(2) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

(3) Now if we integrate (2) in exactly the same manner as (1)

$$(4) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(5) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(6) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(7) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(8) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(9) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

we have the constants  $a$ ,  $b$ , and  $c$  expressed in the desired forms:

only  $V$  remains.

Substituting in

and substituting for

we have

$$(10) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(11) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$

$$(12) \quad \frac{1}{2} \frac{d}{dt} \left( \frac{1}{2} \frac{d^2}{dt^2} \right) = \frac{1}{2} \frac{d^3}{dt^3}$$



$$(16) \text{ In (11) } a = \frac{1}{\sigma_x^2(1-r^2)}$$

$$\text{So } N = \frac{2\pi k \sigma_y}{\sigma_x \sqrt{1-r^2}} = 2\pi k \sigma_x \sigma_y \sqrt{1-r^2}$$

$$(17) \quad k = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}}$$

(18) Now return to the equation

$$Z = k e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

and replace a, h, k, b by their respective values we have

$$Z = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left( \frac{x^2}{\sigma_x^2} - \frac{2xyr}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right)}$$

This is the equation for the correlation surface and r is the coefficient of correlation.

The equation for the normal curve with the x axis as its mean may be shown to be

$$\phi(y) = \frac{N}{\sigma_y \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_y^2}}$$

and that with the y axis as its mean

$$\phi(x) = \frac{N}{\sigma_x \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}}$$

so the probability of the deviation of any y from its mean is

$$\frac{N}{\sigma_y \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_y^2}}$$

and the probability of the deviation of any x from its mean is

$$\frac{N}{\sigma_x \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x^2}}$$

Hence, if these probabilities are independent, the probability of the joint occurrence of these deviations is

$$(19) \quad Z_i = \frac{N^2}{\sigma_x \sigma_y 2\pi} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}$$

which might be considered the equation of a surface.

Now in (18) we showed

$$Z = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left( \frac{x^2}{\sigma_x^2} - \frac{2xyr}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} \right)}$$

$$(12) \quad \frac{1}{N} \sum_{i=1}^N x_i^2 = \frac{1}{N} \sum_{i=1}^N x_i^2$$

$$= \frac{1}{N} \sum_{i=1}^N x_i^2 = \frac{1}{N} \sum_{i=1}^N x_i^2$$

$$(13) \quad \frac{1}{N} \sum_{i=1}^N x_i^2 = \frac{1}{N} \sum_{i=1}^N x_i^2$$

$$= \frac{1}{N} \sum_{i=1}^N x_i^2 = \frac{1}{N} \sum_{i=1}^N x_i^2$$

$$(14) \quad \frac{1}{N} \sum_{i=1}^N x_i^2 = \frac{1}{N} \sum_{i=1}^N x_i^2$$

and replace  $x_i$  by  $y_i$  in the above equation

$$\frac{1}{N} \sum_{i=1}^N y_i^2 = \frac{1}{N} \sum_{i=1}^N y_i^2$$

This is the equation for the correlation surface and  $r$  is the coefficient

of correlation.

The equation for the normal curve with the  $x$  axis as the mean may be

$$(15) \quad \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

and that with the  $y$  axis as the mean

$$(16) \quad \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}$$

is the probability of the deviation of any  $y$  from the mean is

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}$$

and the probability of the deviation of any  $x$  from the mean is

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

hence, if these probabilities are independent, the probability of the

joint occurrence of these deviations is

$$(17) \quad \frac{1}{\sigma^2 2\pi} e^{-\frac{x^2 + y^2}{2\sigma^2}}$$

which may be considered the equation of a surface.

For in (18) we showed

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}$$



was the probability of the joint occurrence of deviations of  $x$  and  $y$ , whether dependent or independent.

If we consider the case where  $N=1$  and  $\mathcal{U}=0$ , then  $\mathcal{Z} = \mathcal{Z}_i$  and we see that when  $\mathcal{U}=0$ ,  $\mathcal{Z} = \mathcal{Z}_i$ . That is when  $\mathcal{U}=0$ ,  $\mathcal{Z}$  is the formula for the probabilities when they are independent. This would suggest that we consider  $r$  a measure for the dependence or correlation of the variables  $x$  and  $y$ .

and the probability of the joint occurrence of  $x$  and  $y$  is then  
 independent or independent.  
 If we consider the case where  $x$  and  $y$  are  $1$  and  $2$ , then we  
 see that when  $x = 1$ ,  $y = 2$  is the only possible for  
 the probabilities when they are independent. This would suggest that we  
 consider a measure for the dependence or correlation of the variables  $x$  and



Chapter V      The Product-Moment Formula for Correlation.

(Jones - "A First Course in Statistics" P. 278)

Since  $\bar{x} = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}$ , and if we

suppose we have  $n$  pairs of associated values of  $x$  and  $y$

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

then any  $x$ , would occur with its  $y$ , in the relation

$$\bar{x} = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x_1^2}{\sigma_x^2} - \frac{2rx_1y_1}{\sigma_x\sigma_y} + \frac{y_1^2}{\sigma_y^2}\right)}$$

The probability that each  $x$  would occur with its associated  $y$ , assuming

the associated pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  were observed independently would be

$$\begin{aligned} \phi(x,y) &= \frac{N}{2\pi\sigma_x\sigma_y(\sqrt{1-r^2})^n} \left\{ e^{-\frac{1}{2(1-r^2)}\left(\frac{x_1^2}{\sigma_x^2} - \frac{2rx_1y_1}{\sigma_x\sigma_y} + \frac{y_1^2}{\sigma_y^2}\right)} \dots e^{-\frac{1}{2(1-r^2)}\left(\frac{x_n^2}{\sigma_x^2} - \frac{2rx_ny_n}{\sigma_x\sigma_y} + \frac{y_n^2}{\sigma_y^2}\right)} \right\} \\ &= \frac{N}{2\pi\sigma_x\sigma_y} \left\{ \frac{1}{(1-r^2)^{\frac{n}{2}}} \cdot e^{-\frac{1}{2(1-r^2)}\left[\frac{\sum x^2}{\sigma_x^2} - \frac{2r\sum(xy)}{\sigma_x\sigma_y} + \frac{\sum y^2}{\sigma_y^2}\right]} \right\} \end{aligned}$$

Let  $K = \frac{\sum(xy)}{N\sigma_x\sigma_y}$  and substitute

$$\phi(x,y) = \frac{N}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2(1-r^2)}\left[\frac{\sum(x^2)}{\sigma_x^2} + \frac{\sum(y^2)}{\sigma_y^2} - 2rnK\right]} \right\}$$

But

$$\frac{\sum(x^2)}{\sigma_x^2} = \frac{\sum(y^2)}{\sigma_y^2} = n$$

$$\phi(x,y) = \frac{N}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2(1-r^2)}[2n - 2rnK]} \right\}$$

$$\phi(x,y) = \frac{n}{2\pi\sigma_x\sigma_y} \left\{ (1-r^2)^{-\frac{n}{2}} \cdot e^{-\frac{n}{(1-r^2)}[1-rK]} \right\}$$

The Product-Moment Formula for Correlation

Chapter V

(1908 - "A First Course in Statistics" - p. 198)

and it is

suppose we have a pair of associated values of  $x$  and  $y$

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

then any  $r$  would occur with the  $y$  in the relation

$$\left( \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \right) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The probability that such a  $r$  would occur with the associated  $y$ , assuming

the associated pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  were observed inde-

pendently would be

$$\left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_x \sigma_y} \exp \left( -\frac{(x_i - \bar{x})^2}{2\sigma_x^2} - \frac{(y_i - \bar{y})^2}{2\sigma_y^2} \right) \right\}$$

for  $x = \bar{x}$  and  $y = \bar{y}$

$$\left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_x \sigma_y} \exp \left( -\frac{(x_i - \bar{x})^2}{2\sigma_x^2} - \frac{(y_i - \bar{y})^2}{2\sigma_y^2} \right) \right\}$$

$$\left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_x \sigma_y} \exp \left( -\frac{(x_i - \bar{x})^2}{2\sigma_x^2} - \frac{(y_i - \bar{y})^2}{2\sigma_y^2} \right) \right\}$$



But  $(1-r^2)^{-\frac{n}{2}} = e^{-\frac{n}{2} \log(1-r^2)}$

$$\therefore f(x,y) = \frac{n}{2\pi\sigma_x\sigma_y} e^{-\frac{n}{2} \log(1-r^2)} \cdot e^{-\frac{n}{1-r^2} (1-rk)}$$

$$f(x,y) = \frac{n}{2\pi\sigma_x\sigma_y} \cdot \frac{1}{e^{\frac{n}{2} \log(1-r^2) + \frac{n}{1-r^2} (1-rk)}}$$

This probability would be the greatest when

$$\frac{1}{2} \log(1-r^2) + \frac{1}{1-r^2} (1-rk) \quad \text{is the least.}$$

Differentiate with respect to  $r$  and equate to 0. This gives the value of  $r$  which will make the expression a minimum and will make the probability the greatest.

$$\frac{1}{2} \cdot \frac{-2r}{1-r^2} + \frac{(1-r^2)(-k) + (1-rk)(2r)}{(1-r^2)^2} = 0$$

$$-r + r^3 + 2r - k - kr^2 = 0$$

$$(r^2 + 1)(r - k) = 0$$

$$r = k$$

The first derivative is  $r^3 - kr^2 + r - k$

The second is  $3r^2 - 2rk + 1$  and when  $r = k$ , this is  $r^2 + 1$ .

Therefore the above expression is a minimum when  $r = k$ .

Hence the probability of the occurrence of the associated pairs of  $x$ 's and  $y$ 's is the greatest when

$$k = r = \frac{\sum(xy)}{N\sigma_x\sigma_y}$$

We assumed that the values of  $x$  and  $y$  were associated and sought a value for  $r$  which would give the maximum probability that for any  $x$  we would obtain its associated  $y$ . This leads to a formula using  $\sum(xy)$  which formula we proceed to develop.

Let  $f(x, y)$  be a function of two variables,  $x$  and  $y$ , which is continuous in a region  $R$  of the  $xy$ -plane. Suppose that  $f$  has partial derivatives  $f_x$  and  $f_y$  in  $R$ . Then the necessary condition for  $f$  to have a local maximum or minimum at a point  $(a, b)$  in  $R$  is that

$$f_x(a, b) = 0 \quad \text{and} \quad f_y(a, b) = 0$$

is necessary but not sufficient. To determine whether  $f$  has a local maximum or minimum at  $(a, b)$ , we must also consider the second partial derivatives of  $f$  at  $(a, b)$ .

$$\Delta = \begin{vmatrix} f_{xx}(a, b) & f_{xy}(a, b) \\ f_{xy}(a, b) & f_{yy}(a, b) \end{vmatrix}$$

If  $\Delta > 0$  and  $f_{xx}(a, b) < 0$ , then  $f$  has a local maximum at  $(a, b)$ .  
 If  $\Delta > 0$  and  $f_{xx}(a, b) > 0$ , then  $f$  has a local minimum at  $(a, b)$ .  
 If  $\Delta < 0$ , then  $f$  has a saddle point at  $(a, b)$ .  
 If  $\Delta = 0$ , the test is inconclusive.

The above test is known as the second derivative test. It is a useful tool for determining the nature of critical points of a function of two variables. However, it is important to note that the test only applies to functions that are twice differentiable at the point in question.

$$f(x, y) = x^2 + y^2$$

is a function of two variables. The partial derivatives are

$$f_x = 2x \quad \text{and} \quad f_y = 2y$$

which are both zero at the origin  $(0, 0)$ . The second partial derivatives are

$$f_{xx} = 2, \quad f_{xy} = 0, \quad \text{and} \quad f_{yy} = 2$$

so that the determinant  $\Delta$  is

$$\Delta = \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0$$

and  $f_{xx}(0, 0) = 2 > 0$ . Therefore,  $f$  has a local minimum at the origin.

It is important to note that the above test only applies to functions that are twice differentiable at the point in question. If the function is not twice differentiable at a point, then the test is inconclusive. In such cases, other methods must be used to determine the nature of the critical point.



The Product Moment Formula

(Forsyth - "Mathematical Analysis of Statistics")

If we define the product-moment of a surface  $z = f(x, y)$  by the relation

$$\bar{z}(x, y) = \iint xy f(x, y) dx dy$$

and if we take the correlation surface

$$z = k e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

then the product moment of the correlation surface about  $\bar{z}$ , the centroid vertical is:

$$(1) \quad \bar{z} xy = k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx dy$$

(2)

$$(3) \quad \text{Now consider only } \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx$$

$$(4) \quad \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx = -\frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy + by) dx$$

$$(5) = -\frac{1}{2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy) dx - \frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx$$

$$(6) \quad \text{Now consider } -\frac{1}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} (-ax - hy) dx$$

(7)

$$(8) = -\frac{1}{a} e^{-\frac{a}{2} \left( \frac{hy^2}{a} - \frac{h^2 y^2}{a^2} \right)} \int_{-\infty}^{\infty} e^{-\frac{a}{2} \left( x^2 + \frac{2hxy}{a} + \frac{h^2 y^2}{a^2} \right)} (-ax - hy) dx$$

$$(9) \quad \text{Now consider only } \int_{-\infty}^{\infty} e^{-\frac{a}{2} \left( x^2 + \frac{2hxy}{a} + \frac{h^2 y^2}{a^2} \right)} (-ax - hy) dx =$$

$$\int_{-\infty}^{\infty} e^{-\frac{a}{2} \left( x + \frac{hy}{a} \right)^2} (-ax - hy) dx = 0$$

It is defined the product-moment of a surface  $Z = f(x, y)$  by the

relation

$$P_{xy} = \iint_R xy \, dA$$

and if we take the correlation surface

$$Z = \frac{xy}{\sigma_x \sigma_y}$$

then the product-moment of the correlation surface about  $x$  and  $y$  is

vertical is:

$$(1) \quad P_{xx} = \iint_R x^2 \, dA$$

$$(2) \quad P_{yy} = \iint_R y^2 \, dA$$

$$(3) \quad P_{xy} = \iint_R xy \, dA$$

$$(4) \quad P_{xx} = \iint_R x^2 \, dA$$

$$(5) \quad P_{yy} = \iint_R y^2 \, dA$$

$$(6) \quad P_{xy} = \iint_R xy \, dA$$

$$(7) \quad P_{xx} = \iint_R x^2 \, dA$$

$$(8) \quad P_{yy} = \iint_R y^2 \, dA$$

$$(9) \quad P_{xy} = \iint_R xy \, dA$$

$$(10) \quad P_{xx} = \iint_R x^2 \, dA$$



(10) Therefore, from (5) and (9)

$$\int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx = -\frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dy$$

$$(11) -\frac{hy}{a} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx = -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2by^2}}$$

shown in section 6 and 7 on page 25.

$$(12) \text{ So } \int_{-\infty}^{\infty} x e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} dx = -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2by^2}}$$

(13) Substituting (12) in (2)

$$\dot{Z}(xy) = h \int_{-\infty}^{\infty} y \cdot -\frac{hy}{a} \sqrt{\frac{2\pi}{a}} e^{-\frac{y^2}{2by^2}} dy$$

(14) Now

$$\dot{Z}(xy) = -\frac{h^2 \sqrt{2\pi}}{a \sqrt{a}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2by^2}} dy$$

$$\int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2by^2}} dy = \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2by^2}} dy \quad \text{integrating by parts.}$$

$$(15) \therefore \dot{Z}(xy) = -\frac{h^2 \sqrt{2\pi}}{a \sqrt{a}} \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2by^2}} dy$$

$$(16) \int_{-\infty}^{\infty} \sigma_y^2 e^{-\frac{y^2}{2by^2}} dy = \sigma_y^3 \sqrt{2\pi}$$

$$(17) \therefore \dot{Z}(xy) = -\frac{2h^2 \sqrt{2\pi}}{a \sqrt{a}} \sigma_y^3$$

(18) In Chapter IV, we found on pages 25 and 26

$$a = \frac{1}{\sigma_x^2 (1-r^2)}$$

$$h = \frac{-r}{\sigma_x \sigma_y (1-r^2)}$$

$$k = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}}$$

$$(19) \text{ So } \dot{Z}(xy) = \frac{-2 \cdot \frac{-r}{\sigma_x \sigma_y (1-r^2)} \cdot \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} \cdot \pi \sigma_y^3}{\frac{1}{\sigma_x^2 (1-r^2)} \cdot \frac{1}{\sigma_x \sqrt{1-r^2}}}$$

$$(20) \dot{Z}(xy) = Nr \sigma_x \sigma_y$$

$$(21) r = \frac{\dot{Z}(xy)}{N \sigma_x \sigma_y}$$





## Chapter VI

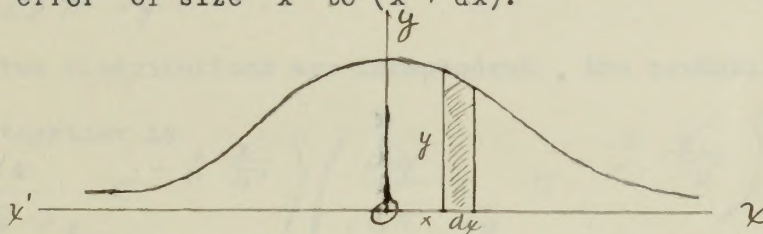
Some Interesting Points Arrived at by  
Considering Normal Correlation.

(Jones - "A First Course in Statistics"  
Chapter XIX )

Frequency Surface for Two Correlated Variables

Before discussing the frequency surface for two correlated variables, it might be helpful to summarize briefly some important features of the normal curve of error.

The equation of the normal curve is  $y = \frac{N}{\sqrt{2\pi} \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$  where  $y dx$  measures the frequency with which a variable deviates from the mean by an amount lying between  $x$  and  $(x+dx)$ ; i.e.  $ydx$  measures the frequency of "error" of size  $x$  to  $(x+dx)$ .



The probability of an error lying between  $x_1$  and  $x_2$  is given by the ratio.

$$\frac{\text{frequency of all errors between the given limits}}{\text{frequency of all errors}}$$

The frequency of all errors is  $\int_{-\infty}^{\infty} y dx = N$

Therefore, the probability of frequency of errors between  $x$  and  $x+dx$  is  $\frac{y dx}{N}$

$$= \frac{dx}{\sqrt{2\pi} \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$$

(Notes - "Short Course in Statistics"  
 Chapter III)

Probability Distributions for Two Correlated Variables

Before discussing the bivariate normal distribution for two correlated variables, we shall be obliged to assume that the two variables are jointly normally distributed. The bivariate normal distribution is defined as follows:

The probability density function of the bivariate normal distribution is given by the following expression:  $f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x-\mu_x}{\sigma_x} - \rho\frac{y-\mu_y}{\sigma_y}\right]^2 - \frac{1}{2}\frac{(y-\mu_y)^2}{\sigma_y^2}\left[1-\rho^2\right]\right\}$  where  $\mu_x, \mu_y$  are the means,  $\sigma_x, \sigma_y$  are the standard deviations, and  $\rho$  is the correlation coefficient. The probability of an error between  $x$  and  $y$  is given by the following expression:



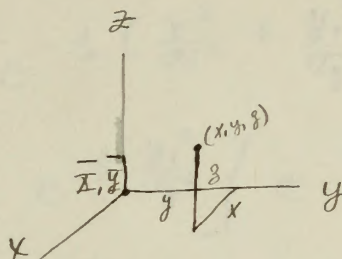
The probability of an error between  $x$  and  $y$  is given by the following expression:  $P(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$ . The probability of all errors is given by the following expression:  $P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$ .

The probability of all errors is given by the following expression:  $P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$ . Therefore, the probability of an error between  $x$  and  $y$  is given by the following expression:  $P(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$ .



Frequency Surface, showing the distribution of two completely independent variables each subject to the normal law.

If  $\bar{X}$ ,  $\bar{y}$  be taken as the origin and  $x$  and  $y$  the deviations from  $\bar{X}$  &  $\bar{y}$  respectively,



then the probability of a deviation between  $x$  and  $x + dx$  is

$$\frac{dx}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \frac{x^2}{\sigma_x^2}}$$

and the probability of a deviation between  $y$  and  $y + dy$  is

$$\frac{dy}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2} \frac{y^2}{\sigma_y^2}}$$

Since the two distributions are independent, the probability of their occurring together is

$$\begin{aligned} & \left( \frac{dx}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \frac{x^2}{\sigma_x^2}} \right) \left( \frac{dy}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2} \frac{y^2}{\sigma_y^2}} \right) \\ &= \frac{dx dy}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)} \end{aligned}$$

If  $N$  is the total number of observations, the frequency with which such deviations occur together is

$$\frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)} dx dy$$

$$\text{If } \oint dx dy = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)} dx dy$$

$$\text{then } \oint = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}$$

frequency distribution, assuming the distribution of two normally independent

variables each subject to the normal law.

Let  $X, Y$  be taken as the origin and  $x$  and  $y$  the deviation from

$X = \bar{x}$  respectively,



then the probability of a deviation between  $x$  and  $x + dx$  is

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx$$

and the probability of a deviation between  $y$  and  $y + dy$  is

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} dy$$

Since the two distributions are independent, the probability of their

occurring together is

$$\left( \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \right) \left( \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} \right) = \frac{1}{\sigma^2 2\pi} e^{-\frac{x^2 + y^2}{2\sigma^2}}$$

If  $N$  is the total number of observations, the frequency with which

and deviations occur together is

$$N \cdot \frac{1}{\sigma^2 2\pi} e^{-\frac{x^2 + y^2}{2\sigma^2}} = \frac{N}{\sigma^2 2\pi} e^{-\frac{x^2 + y^2}{2\sigma^2}}$$



This, then, is the equation of the frequency surface when the variables are independent.

Now let us discuss this surface to see what it is like.

If  $y = y_1$ ,

$$\begin{aligned} Z &= \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y_1^2}{\sigma_y^2}\right)} \\ &= \left[ \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\frac{y_1^2}{\sigma_y^2}} \right] e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \\ &= \frac{N_1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \quad \text{where } N_1 = \frac{N}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\frac{y_1^2}{\sigma_y^2}} \end{aligned}$$

$$\text{But } Z = \frac{N_1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}}$$

is the equation of a normal curve

in which  $\bar{x}$ ,  $\sigma_x$  are not affected by different values of  $y$ . Hence all arrays of  $\bar{x}$  are similar, having the same mean and the same standard deviation.

Since the surface is symmetrical, the same may be said for all the arrays of  $\bar{y}$ .

Furthermore, if  $Z = k$ , a constant

$$\begin{aligned} k &= \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \\ \frac{2\pi k\sigma_x\sigma_y}{N} &= e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \end{aligned}$$

Since the left-hand side is a constant, the right-hand side is also,

so

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = C, \quad \text{a constant.}$$

That is the equation of an ellipse; thus we may say where  $x$  and  $y$  occur with the frequency  $k$ , the points  $(x, y)$  lie on an ellipse in the plane  $Z = k$  or

$$\begin{cases} Z = k \\ \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = C \end{cases} \quad \text{defines the locus of}$$

This, then, is the equation of the trajectory surface when the particles

are identical.

Now let us discuss this surface in case  $\mu \neq 1$ .

$$1 - \frac{1}{2} \left( \frac{y}{x} + \frac{y_1}{x_1} \right)$$

$$\left[ \frac{y_1}{x_1} - \frac{y}{x} \right] = \frac{1}{2} \left( \frac{y_1}{x_1} + \frac{y}{x} \right)$$

$$\frac{y_1}{x_1} - \frac{y}{x} = \frac{1}{2} \left( \frac{y_1}{x_1} + \frac{y}{x} \right)$$

is the equation of a normal curve

in which  $x, y$  are not affected by different values of  $y_1$ . Hence

all curves of  $x$  are alike, having the same form and the same dimensions

each time.

When the surface is considered, the same may be said for all the

curves of  $y$ .

Furthermore, if  $\mu$  is a constant

$$\frac{y_1}{x_1} - \frac{y}{x} = \frac{1}{2} \left( \frac{y_1}{x_1} + \frac{y}{x} \right)$$

$$\frac{y_1}{x_1} - \frac{y}{x} = \frac{1}{2} \left( \frac{y_1}{x_1} + \frac{y}{x} \right)$$

Since the left-hand side is a constant, the right-hand side is also.

$$\frac{y_1}{x_1} + \frac{y}{x} = \text{a constant.}$$

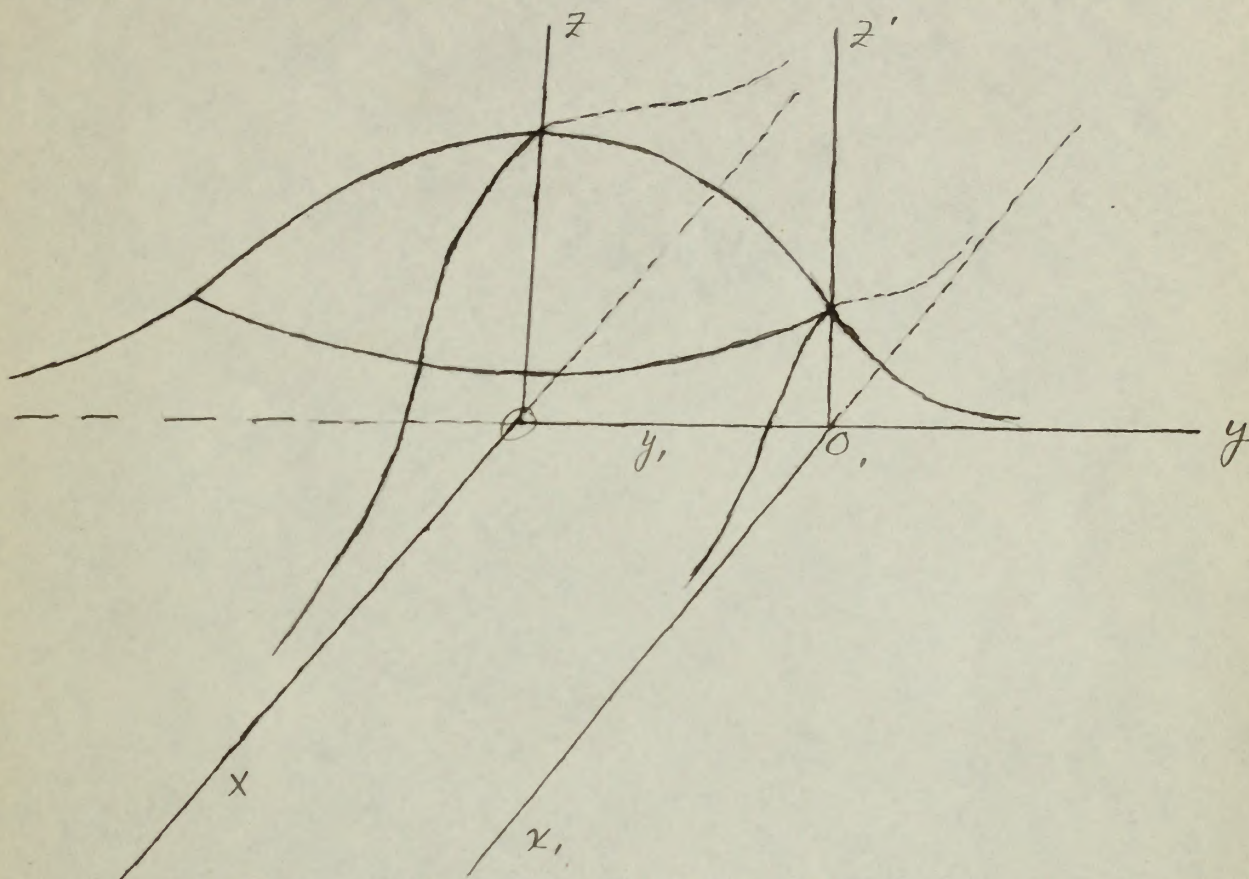
But is the equation of an ellipse, that is, any curve where  $x$  and  $y$  occur

with the frequency  $\mu$ , the point  $(x, y)$  lies on an ellipse in the  $xy$ -plane.

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$



points where  $x$  and  $y$  occur with the same frequency. If we vary  $k$ , the frequency, and consequently vary  $c$ , we get a series of ellipses. If we project these orthogonally on the plane  $\bar{X} O \bar{Y}$  we would get a series of concentric similar ellipses. This enables us to draw the surface



Frequency Surface for Two Correlated Variables.

If we consider the variables  $\bar{X}$  and  $\bar{Y}$  and take  $\bar{X}, \bar{Y}$  as the origin so that  $\bar{X} - \bar{X} = x$  and  $\bar{Y} - \bar{Y} = y$ , then the line of regression giving the best  $y$  corresponding to any  $x$  is  $y = r \frac{\sigma_y}{\sigma_x} x$

If we consider  $\eta$  the error made in taking  $y$  from this equation instead of the observed  $y$ , then  $\eta = y - r \frac{\sigma_y}{\sigma_x} x$ . Thus for every  $(x, y)$  there is an  $\eta$  and the same  $\eta$  occurs as often as any pair  $(x, y)$  is repeated; thus the frequency distribution  $(x, \eta)$  is exactly the same as that of  $(x, y)$ . The correlation between  $\eta$  and  $x$  is  $\frac{\sum (x\eta)}{N \sigma_x \sigma_\eta}$  and should equal zero.





$$\begin{aligned}
\sum(xn) &= \sum \left[ x \left( y - r \frac{\sigma_y}{\sigma_x} x \right) \right] \\
&= \sum(xy) - r \frac{\sigma_y}{\sigma_x} \sum(x^2) \\
&= NP - \frac{P}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} \cdot N \sigma_x^2 \\
&= NP - NP \\
\sum(xn) &= 0
\end{aligned}$$

Therefore, the variables  $x$  and  $n$  are independent and the probability of their occurring together is the product of their separate probabilities.

The probability of a deviation between  $x$  and  $(x+dx)$  occurring if we consider  $x$  alone is

$$\frac{dx}{\sqrt{2\pi} \sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}$$

and the probability of a deviation between  $n$  and  $(n+dn)$  is

$$\frac{dn}{\sqrt{2\pi} \sigma_n} e^{-\frac{n^2}{2\sigma_n^2}}$$

The probability of a combined occurrence of these deviations is

$$\begin{aligned}
&\frac{dx dn}{2\pi \sigma_x \sigma_n} e^{-\frac{1}{2} \left\{ \frac{x^2}{\sigma_x^2} + \frac{n^2}{\sigma_n^2} \right\}} \\
(1) \quad &= \frac{dx dn}{2\pi \sigma_x \sigma_n} e^{-\frac{1}{2} \left\{ \frac{x^2}{\sigma_x^2} + \frac{\left( y - r \frac{\sigma_y}{\sigma_x} x \right)^2}{\sigma_n^2} \right\}} \\
(2) \quad &= \frac{dx dn}{2\pi \sigma_x \sigma_n} e^{-\frac{1}{2} \left[ \frac{y^2}{\sigma_n^2} - \frac{2xyr\sigma_x}{\sigma_x \sigma_n^2} + x^2 \left( \frac{1}{\sigma_x^2} + \frac{r^2 \sigma_y^2}{\sigma_x^2 \sigma_n^2} \right) \right]}
\end{aligned}$$

$$\begin{aligned}
\text{But } N \sigma_n^2 &= \sum \left( y - r \frac{\sigma_y}{\sigma_x} x \right)^2 \\
&= \sum y^2 - 2r \frac{\sigma_y}{\sigma_x} \sum(xy) + \frac{r^2 \sigma_y^2}{\sigma_x^2} \sum(x^2) \\
&= N \sigma_y^2 + N r^2 \sigma_y^2 \\
(3) \quad &= N \sigma_y^2 (1 - r^2)
\end{aligned}$$

Similarly if  $g$  is the error made in estimating any  $x$  from  $X = r \frac{\sigma_x}{\sigma_y} y$  then

$$(4) \quad N \sigma_g^2 = N \sigma_x^2 (1 - r^2)$$





$$(5) \quad \frac{\sigma_g^2}{\sigma_x^2} = 1 - r^2 = \frac{\sigma_n^2}{\sigma_y^2}$$

$$(6) \quad \text{So } \frac{\sigma_y}{\sigma_x \sigma_n} = \frac{1}{\sigma_x \sigma_n} \cdot \frac{\sigma_y}{\sigma_n} = \frac{1}{\sigma_x \sigma_n \sigma_g} = \frac{1}{\sigma_n \sigma_g}$$

$$(7) \quad \text{Also } \frac{1}{\sigma_x^2} + \frac{r^2 \sigma_y^2}{\sigma_x^2 \sigma_n^2} = \frac{1}{\sigma_x^2} \left( 1 + r^2 \frac{\sigma_y^2}{\sigma_n^2} \right) \\ = \frac{1}{\sigma_x^2} \left( 1 + r^2 \cdot \frac{1}{1-r^2} \right) \\ = \frac{1}{\sigma_x^2} \cdot \frac{1}{1-r^2} \\ = \frac{1}{\sigma_g^2}$$

$$(8) \quad \text{Substituting (3), (4), (6) and (7) in (2)} \\ = \frac{dx dn}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left( \frac{y^2}{\sigma_n^2} - 2xyr \frac{1}{\sigma_g \sigma_n} + \frac{x^2}{\sigma_g^2} \right)}$$

This is the probability of the combined occurrence of deviations  $x$  to  $(x+dx)$ ,  $n$  to  $(n+dn)$ . Now if we substitute (3) and (4) in (8) we get the frequency of the combined occurrence of deviations  $x$  to  $(x+dx)$  and  $y$  to  $(y+dy)$

$$(9) = \frac{dx dy}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left( \frac{y^2}{\sigma_y^2 (1-r^2)} - 2xyr \frac{1}{\sigma_x \sigma_y (1-r^2)} + \frac{x^2}{\sigma_x^2 (1-r^2)} \right)} \\ = \frac{dx dy}{2\pi \sigma_x \sigma_y \sqrt{1-r^2}} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}}$$

Thus if  $\mathcal{F} dx dy$  represents this frequency where  $N$  is the total number of observations

$$\mathcal{F} = \frac{N}{2\pi \sqrt{1-r^2} \sigma_x \sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}}$$

This equation represents the frequency surface for two correlated variables.

$$\begin{aligned}
 (a) \quad & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
 & = \frac{1}{2} \begin{pmatrix} 1+1 & 1-1 \\ 1+1 & 1-1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \\
 (b) \quad & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
 & = \frac{1}{2} \begin{pmatrix} 1+1 & 1-1 \\ 1+1 & 1-1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}
 \end{aligned}$$

$$\begin{aligned}
 (c) \quad & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
 & = \frac{1}{2} \begin{pmatrix} 1+1 & 1-1 \\ 1+1 & 1-1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}
 \end{aligned}$$

This is the probability of the combined occurrence of both events.

$$\begin{aligned}
 (d) \quad & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
 & = \frac{1}{2} \begin{pmatrix} 1+1 & 1-1 \\ 1+1 & 1-1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}
 \end{aligned}$$

$$\begin{aligned}
 (e) \quad & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
 & = \frac{1}{2} \begin{pmatrix} 1+1 & 1-1 \\ 1+1 & 1-1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}
 \end{aligned}$$

This relation represents the frequency surface for two correlated variables.



Let us now study this equation in order to learn some of its interesting points.

$$\text{Let } t \text{ (a constant)} = \frac{N}{2\pi \sqrt{1-r^2} \sigma_x \sigma_y}$$

and consider the surface

$$Z = t e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}}$$

If we let  $y = y_1$ , we have

$$\begin{aligned} Z &= t e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y_1^2}{\sigma_y^2} - 2r \frac{xy_1}{\sigma_x \sigma_y} \right) \frac{1}{1-r^2}} \\ &= t e^{-\frac{1}{2(1-r^2)} \left\{ \frac{y_1^2}{\sigma_y^2} (1-r^2) + \left( \frac{x}{\sigma_x} - r \frac{y_1}{\sigma_y} \right)^2 \right\}} \\ &= t \cdot e^{-\frac{y_1^2}{2\sigma_y^2}} \cdot e^{-\frac{1}{2(1-r^2)} \left( \frac{x}{\sigma_x} - r \frac{y_1}{\sigma_y} \right)^2} \end{aligned}$$

$$(1) \quad Z = \frac{t}{C \frac{y_1^2}{\sigma_y^2}} \cdot e^{-\frac{1}{2\sigma_x^2(1-r^2)} \left( x - r y_1 \frac{\sigma_x}{\sigma_y} \right)^2}$$

This is the equation arrived at by taking the equation of the normal curve  $Z = \frac{t}{C \frac{y_1^2}{\sigma_y^2}} e^{-\frac{1}{2\sigma_x^2(1-r^2)} x^2}$ , an equation in  $x$  and  $z$  in the plane  $y = y_1$ , and shifting through a distance  $r y_1 \frac{\sigma_x}{\sigma_y}$  along an axis parallel to  $OX$ . The equation is that of a normal curve with a standard deviation  $\sigma_x \sqrt{1-r^2}$  and the mean at the intersection of the planes  $y = y_1$  and  $x = r y_1 \frac{\sigma_x}{\sigma_y}$ . So the greatest frequency in this particular distribution is  $Z = \frac{t}{C \frac{y_1^2}{\sigma_y^2}}$ , determined by the intersection of the two planes above.

$$\text{If } y = 0 \\ Z = t e^{-\frac{1}{2(1-r^2)} \frac{x^2}{\sigma_x^2}}$$

This is a normal curve with a standard deviation  $\sigma_x \sqrt{1-r^2}$  and the mean at the origin where  $Z = t$ . This mean may be considered the intersection of the plane  $y = 0$  and  $\frac{x}{\sigma_x} = r \frac{y}{\sigma_y}$ .

Thus the planes giving the means of the  $x$ 's corresponding to particular

Let us now study this equation in order to learn some of its interesting

features.

1.

$$\text{Let } F(x) = \int_{-\infty}^x f(t) dt$$

$$\text{and consider the function } F(x) = \int_{-\infty}^x f(t) dt$$

$$\text{If we let } x = \frac{y}{\sigma} \text{ we have } F(x) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$\left\{ \begin{aligned} &F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt \\ &F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt \end{aligned} \right.$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

This is the function which is used to define the equation of the normal

curve. It is a function of  $x$  and  $y$  in the plane.

It is a function of  $x$  and  $y$  in the plane. It is a function of  $x$  and  $y$  in the plane.

The equation is that of a normal curve with a standard deviation

of  $\sigma$ . The mean of the distribution of the values  $y = \sigma x$  is

$\mu = \sigma x$ . In the present case,  $\mu = 0$ . In this particular distribution

the mean is  $\mu = 0$ . The standard deviation is  $\sigma = 1$ .

$$F\left(\frac{y}{\sigma}\right) = \int_{-\infty}^{\frac{y}{\sigma}} f(t) dt$$

This is a normal curve with a standard deviation of  $\sigma = 1$  and a mean of

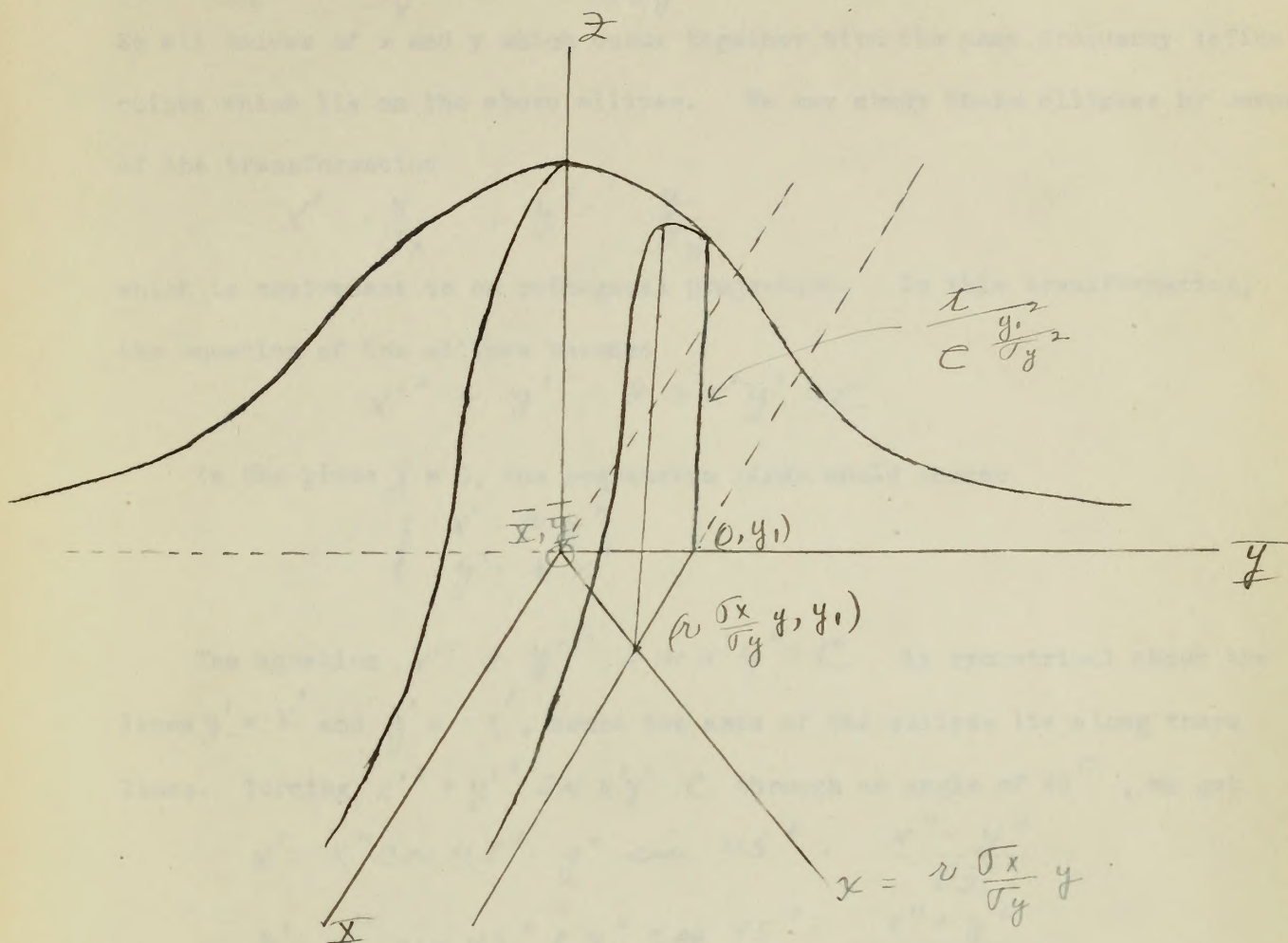
$\mu = 0$ . This curve may be considered the distribution of

$$\text{the value } y = \sigma x = \frac{y}{\sigma}$$

from the plane. It is the mean of the  $x$ 's corresponding to a particular



values of  $y$  meet  $\bar{z} = 0$  in the regression line  $\frac{x}{\sigma_x} = r \frac{y}{\sigma_y}$ .



Thus the  $x$  arrays all have the same standard deviation  $\sigma_x \sqrt{1-r^2}$ , and all have their means at the intersections of the planes through particular values of  $y$  and the plane through the lines  $\bar{z} = 0$  and the regression line

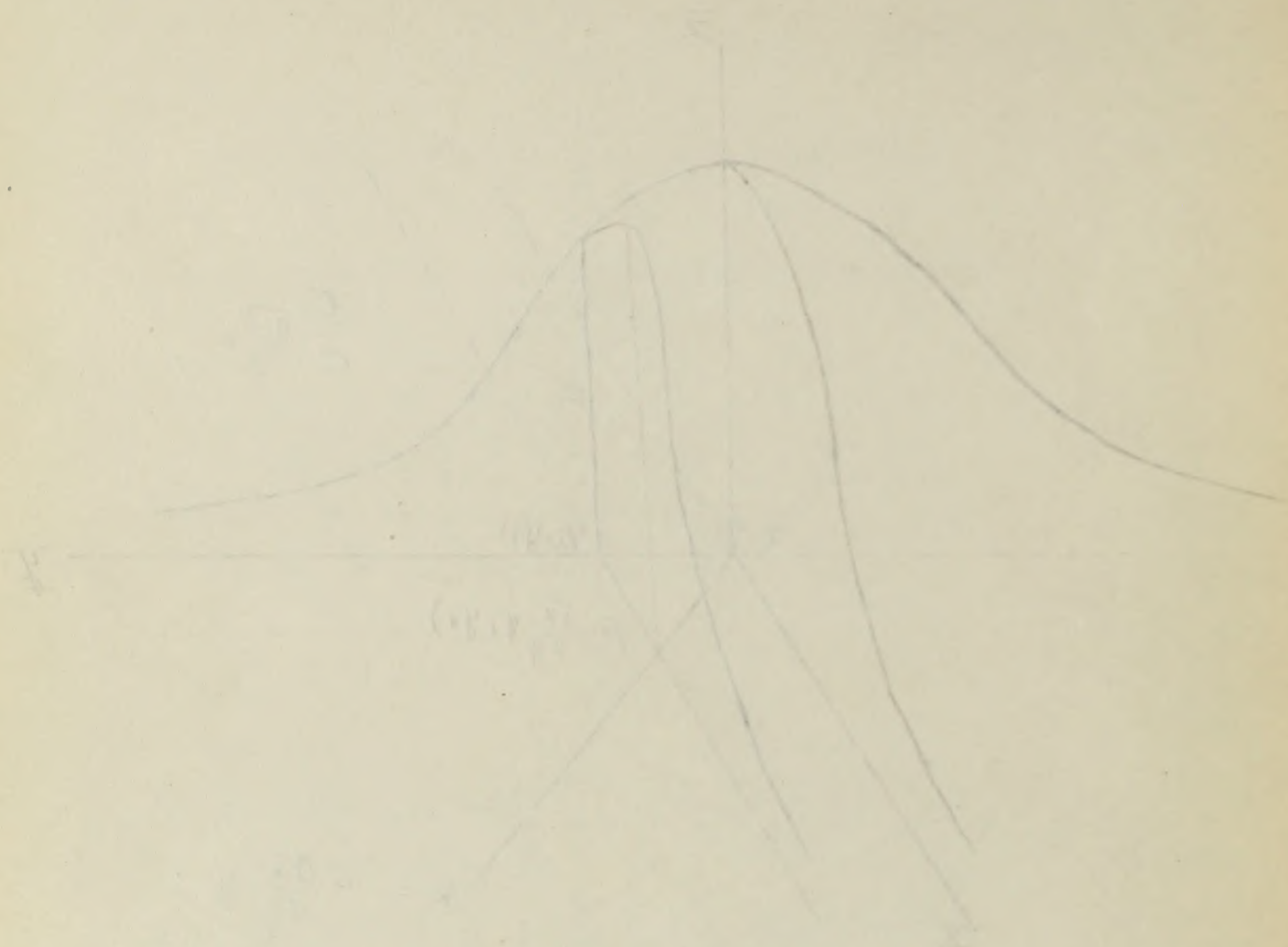
$$x = r \frac{\sigma_x}{\sigma_y} y.$$

Similarly it can be shown that all the  $y$  arrays are normal distributions, have the same standard deviation  $\sigma_y \sqrt{1-r^2}$  and have their means at the intersection of the planes through particular values of  $x$  and the plane determined by the lines  $\bar{z} = 0$  and the regression line  $y = r \frac{\sigma_y}{\sigma_x} x$ .

If we consider the equation

$$\bar{z} = t e^{-\frac{t^2}{2(1-r^2)}} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} \right)$$

values of  $y$  mean  $x = 0$  in the regression line.



Thus the  $x$  curves all have the same standard deviation  $\sigma_x^2$  and all have their means at the intersections of the planes through particular values of  $y$  and the plane through the line  $y = 0$  and the regression line.

Similarly it can be shown that all the  $y$  curves are normal distributions, have the same standard deviation  $\sigma_y^2$ , and have their means at the intersections of the planes through particular values of  $x$  and the plane determined

by the lines  $y = 0$  and the regression line  $y = \frac{\sigma_{xy}}{\sigma_x^2} x$ .

If we consider the equation  $\frac{\sigma_{xy}}{\sigma_x^2} x = y$  we can see that the regression line is the line of intersection of the planes through particular values of  $x$  and the plane determined



and let  $\mathcal{A}$  equal a constant,  $k$ , then

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r \frac{xy}{\sigma_x \sigma_y} = C, \text{ a constant.}$$

So all values of  $x$  and  $y$  which occur together with the same frequency define points which lie on the above ellipse. We may study these ellipses by means of the transformation

$$x' = \frac{x}{\sigma_x}, \quad y' = \frac{y}{\sigma_y}$$

which is equivalent to an orthogonal projection. In this transformation, the equation of the ellipse becomes

$$x'^2 + y'^2 - 2r x' y' = C$$

In the plane  $z = 0$ , the regression lines would become

$$\begin{cases} x' = r y' \\ y' = r x' \end{cases}$$

The equation  $x'^2 + y'^2 - 2r x' y' = C$  is symmetrical about the lines  $y' = x'$  and  $y' = -x'$ , hence the axes of the ellipse lie along these lines. Turning  $x'^2 + y'^2 - 2r x' y' = C$  through an angle of  $45^\circ$ , we get

$$x' = x'' \cos 45^\circ - y'' \sin 45^\circ = \frac{x'' - y''}{\sqrt{2}}$$

$$y' = x'' \sin 45^\circ + y'' \cos 45^\circ = \frac{x'' + y''}{\sqrt{2}}$$

so the equation becomes

$$\frac{(x'' - y'')^2}{2} + \frac{(x'' + y'')^2}{2} - 2r \frac{(x'' - y'')(x'' + y'')}{2} = C$$

$$x''^2 + y''^2 - r(x''^2 - y''^2) = C$$

$$\frac{x''^2}{\frac{C}{1-r}} + \frac{y''^2}{\frac{C}{1+r}} = 1$$

Hence the semi-major axis is  $\sqrt{\frac{C}{1-r}}$  and the semi-minor axis is  $\sqrt{\frac{C}{1+r}}$ . As  $r$  increases from 0 to 1, the semi-major axis increases from  $\sqrt{C}$  to  $\infty$  and the semi-minor axis decreases from  $\sqrt{C}$  to  $\sqrt{\frac{C}{2}}$ ; as  $r$  decreases from 0 to -1, the semi-major axis decreases from  $\sqrt{C}$  to  $\sqrt{\frac{C}{2}}$

and let  $\lambda$  be a constant,  $\lambda \neq 0$ .

$$C = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{is a constant.}$$

To all values of  $x$  and  $y$  which occur together in the same frequency distribution, we assign the same value of  $z$ . The new study these ellipses by means of the transformation

$$x' = \frac{x+y}{\sqrt{2}}, \quad y' = \frac{x-y}{\sqrt{2}}$$

which is equivalent to an orthogonal rotation. In this transformation,

the equation of the ellipse becomes

$$x'^2 + y'^2 = 2 + x'y' - C$$

In the plane  $z = 1$ , the regression lines would be

$$\begin{cases} x' = y' \\ y' = x' \end{cases}$$

The regression  $x' = y'$  is represented above the

line  $y' = x'$ , since the axis of the ellipse is at an angle of  $45^\circ$  to the

line. In the plane  $z = 0$ , the regression lines would be

$$x' = y' \quad \text{and} \quad y' = x'$$

$$x' = y' \quad \text{and} \quad y' = x'$$

to the coordinate planes

$$\frac{(x'' - y'')^2}{2} + \frac{(x'' + y'')^2}{2} = 2$$

$$x''^2 + y''^2 = 2$$

$$\frac{x''^2}{2} + \frac{y''^2}{2} = 1$$

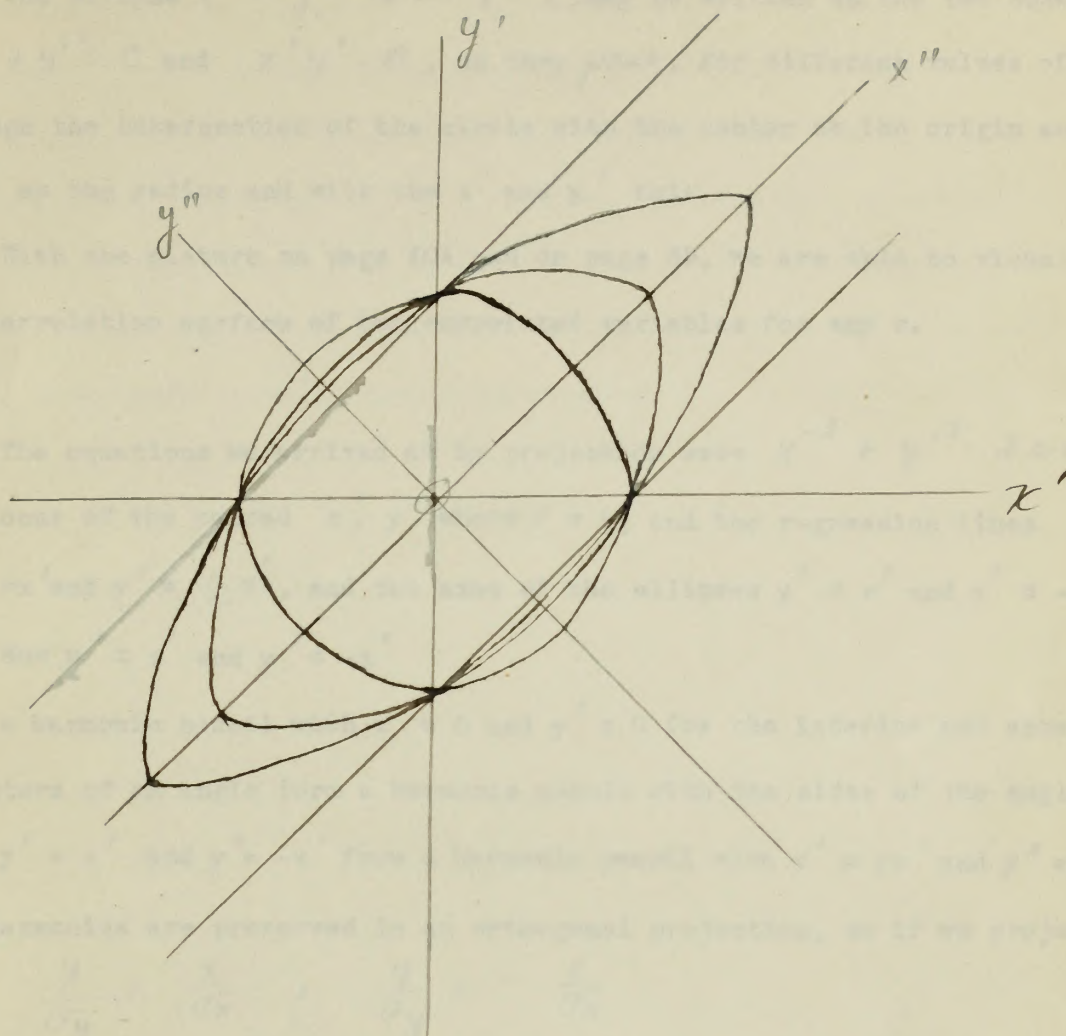
and the semi-minor axis is

as  $z$  increases from 0 to 1, the semi-minor axis increases

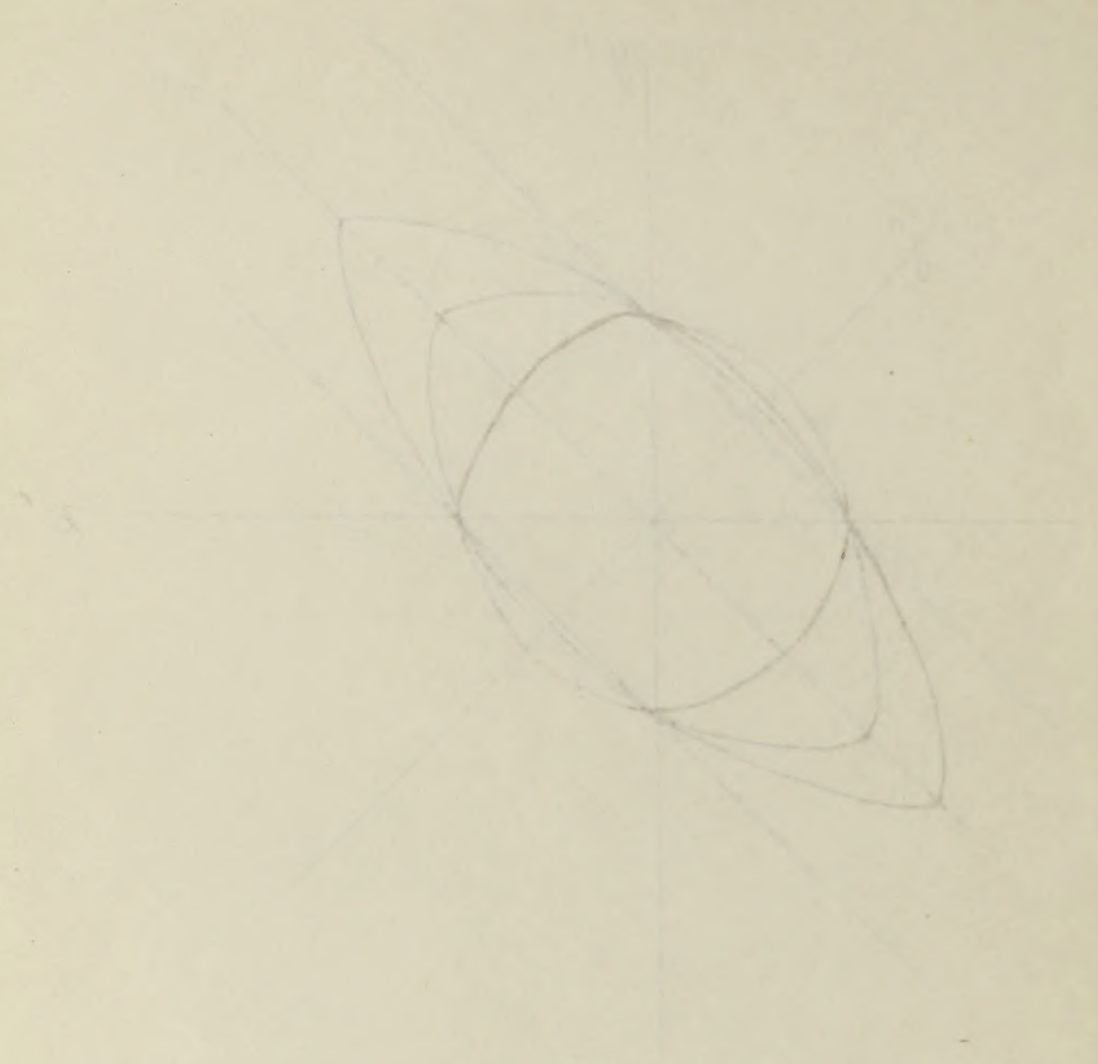
from 0 to 1, and the semi-major axis decreases from  $\sqrt{2}$  to 1

as  $z$  increases from 0 to 1, the semi-major axis decreases from  $\sqrt{2}$  to 1





Showing projections in the  $x y$  plane of the plane  $z = k$   
 with correlation surfaces having different values of  $r$ .



Showing projections in the  $x, y$  plane of the sphere  $x^2 + y^2 + z^2 = 1$   
with corresponding surfaces having different values of  $c$ .



and the semi-minor axis increases from  $\sqrt{C}$  to  $\infty$ .

The ellipse  $x'^2 + y'^2 - 2rx'y' = C$  may be written as the two equations  $x'^2 + y'^2 = C$  and  $x'y' = 0$ , so they pass, for different values of  $C$ , through the intersection of the circle with the center at the origin and  $\sqrt{C}$  as the radius and with the  $x'$  and  $y'$  axis.

With the picture on page 40A and on page 39, we are able to visualize the correlation surface of two correlated variables for any  $r$ .

The equations we arrived at by projection were  $x'^2 + y'^2 - 2rx'y' = C$ , the locus of the paired  $x', y'$  where  $\lambda = k$ , and the regression lines  $y' = rx'$  and  $y' = \frac{1}{r}x'$ , and the axes of the ellipses  $y' = x'$  and  $y' = -x'$ .

Now  $y' = x'$  and  $y' = -x'$

form a harmonic pencil with  $x' = 0$  and  $y' = 0$  for the interior and exterior bisectors of an angle form a harmonic pencil with the sides of the angle.

Also  $y' = x'$  and  $y' = -x'$  form a harmonic pencil with  $y' = rx'$  and  $y' = \frac{1}{r}x'$ .

Now harmonics are preserved in an orthogonal projection, so if we project back,

$$\frac{y}{\sigma_y} = \frac{x}{\sigma_x}, \quad \frac{y}{\sigma_y} = -\frac{x}{\sigma_x}$$

are harmonic with

$$x = 0, \quad y = 0$$

and

$$\frac{y}{\sigma_y} = \frac{x}{\sigma_x}, \quad \frac{y}{\sigma_y} = -\frac{x}{\sigma_x}$$

are harmonic with

$$\frac{y}{\sigma_y} = r \frac{x}{\sigma_x} \quad \text{and} \quad \frac{y}{\sigma_y} = \frac{1}{r} \frac{x}{\sigma_x}$$

Thus we say the two lines of regression corresponding to maximum correlation ( $r = +1, r = -1$ ) are harmonic with

(1) The axes,

(2) The lines of regression for any  $r$ .

but the real-orthogonal axes distances from  $O$  to  $C$  and  $O$  to  $C'$  are

The ellipses  $C$  and  $C'$  may be written as the two equations

$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$  and  $\frac{x'^2}{a'^2} + \frac{y'^2}{b'^2} = 1$ , so that  $C$  and  $C'$  are different ellipses of  $C$

through the intersection of the ellipse with the center of the origin and

$C$  as the center and with the  $x$  and  $y$  axes.

With the ellipse in page 401 and on page 30, we are able to visualize

the correlation curves of two correlated variables for any  $r$ .

The equations are derived as by projection were  $y' = y \cos \theta + x \sin \theta$

the locus of the point  $(x', y')$  where  $x = 1$ , and the regression lines

$y' = r x'$  and  $x' = r y'$ , and the axes of the ellipse  $y' = 1$  and  $x' = 1$

are  $y' = 1$  and  $x' = 1$

Form a harmonic pencil with  $x' = 0$  and  $y' = 0$  for the intersection and exterior

divisors of an angle form a harmonic pencil with the sides of the angle.

Since  $y' = 1$  and  $x' = 1$  form a harmonic pencil with  $x' = 0$  and  $y' = 0$

the harmonic and projected in an orthogonal projection, so if we project both

are harmonic with

and

are harmonic with

and

are harmonic with

and

are harmonic with

and

are harmonic with

Thus we see the two lines of regression corresponding to certain corre-

lation (e.g.  $r = 1$ ,  $r = -1$ ) are harmonic with

(1) The axes,

(2) The lines of regression for any  $r$ .



In other words, the lines of regression corresponding to maximum correlation bisect the interior and exterior angles formed by the lines of regression for any  $r$ , a fact which we have proved in a previous section.

We have shown that in an ideal distribution the means of the rows and the means of the columns all lie on the regression lines and in our previous work we have generalized this by assuming that the distributions we have had were so chosen that if there were an infinite number of paired values of  $x$  and  $y$ , they would form an ideal distribution.

In other words, the lines of regression corresponding to various combinations of the independent and extraneous variables formed by the lines of regression for any  $x$ , a fact which we have proved in a previous section.

We have shown that in an ideal distribution the means of the  $y$ 's and the means of the  $x$ 's are the same. It is to the regression lines and is not necessary that we have proved that the distribution of the  $y$ 's is normal and that the distribution of the  $x$ 's is normal. It is to the regression lines and is not necessary that we have proved that the distribution of the  $y$ 's is normal and that the distribution of the  $x$ 's is normal.



## Chapter VII

Computation Formulas for the Coefficient of  
Correlation. Problems.

Having developed the formula

$$r = \frac{\sum_{i=1}^n x_i y_i}{N \sigma_x \sigma_y}$$

where

$$x_i = \bar{X}_i - \bar{X}$$

$$y_i = \bar{Y}_i - \bar{Y}$$

we may rewrite this in several ways which will aid in the numerical computation.

$$(1) \quad r = \frac{\sum_{i=1}^n x_i y_i}{N \sigma_x \sigma_y}$$

Formula I

Origin at the true mean.

$$(2) \quad r = \frac{\frac{1}{N} \sum x y}{\sqrt{\frac{1}{N} \sum x^2} \sqrt{\frac{1}{N} \sum y^2}}$$

by substituting  $\sigma_x = \sqrt{\frac{1}{N} \sum x^2}$ ,  $\sigma_y = \sqrt{\frac{1}{N} \sum y^2}$

$$\text{so } r = \frac{\sum x y}{\sqrt{\sum x^2} \sqrt{\sum y^2}}$$

Formula II

Origin at the true mean.

$$(3) \quad r = \frac{\frac{1}{N} \sum (\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y})}{\sigma_x \sigma_y}$$

where

$$x_i = \bar{X}_i - \bar{X}$$

$$y_i = \bar{Y}_i - \bar{Y}$$

Formula III

Origin at True Mean.

(4) Now to rewrite this formula so that the deviations will refer to some point other than the true mean as origin.

Derivation of the Formulas

$$\frac{1}{\lambda} = \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \dots$$

$$\frac{1}{\lambda} = \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \dots$$

As we have seen in several ways which will be in the Appendix

Correlation

Formula I

Origin of the time zero

$$\frac{1}{\lambda} = \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \dots$$

(1)

Formula II

Origin of the time zero

$$\frac{1}{\lambda} = \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \dots$$

so

(2)

Formula III

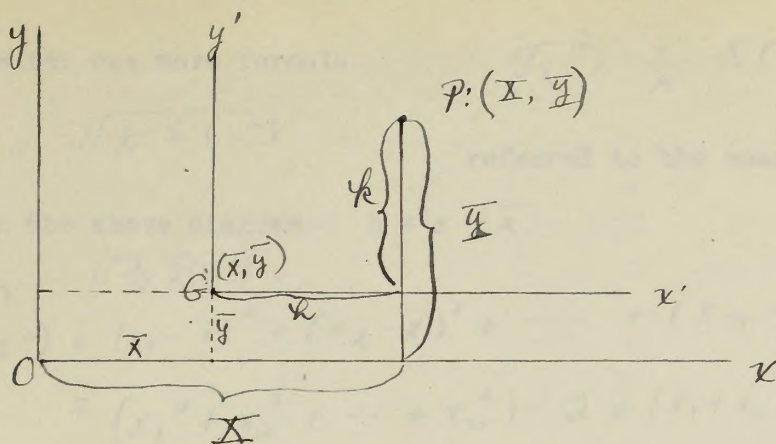
Origin of the time zero

$$\frac{1}{\lambda} = \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \dots$$

(3) As we have seen this formula is that the deviations will refer to

some point other than the time zero as origin





Let  $h = x - \bar{x}$

$k = y - \bar{y}$

In the formula  $r = \frac{\sum x_i y_i}{N \sigma_x \sigma_y}$   $x$  and  $y$  represented the deviations from the mean, and in the above  $h$  and  $k$  are the respective deviations, so we may rewrite

$$r = \frac{\sum h k}{N \sigma_x \sigma_y}$$

$$\begin{aligned} \sum (h k) &= (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \\ &= (x_1 y_1 - \bar{x} y_1 - x_1 \bar{y} + \bar{x} \bar{y}) + \dots + (x_n y_n - \bar{x} y_n - x_n \bar{y} + \bar{x} \bar{y}) \\ &= (x_1 y_1 + x_2 y_2 + \dots + x_n y_n) - \bar{x} (y_1 + y_2 + \dots + y_n) - \bar{y} (x_1 + x_2 + \dots + x_n) + n \bar{x} \bar{y} \\ &= \sum (x_i y_i) - \bar{x} \sum y_i - \bar{y} \sum x_i + n \bar{x} \bar{y} \end{aligned}$$

But  $\frac{\sum y_i}{n} = \bar{y}$  and  $\frac{\sum x_i}{n} = \bar{x}$

$$\therefore \sum (h k) = \sum (x_i y_i) - \bar{x} n \bar{y} - \bar{y} n \bar{x} + n \bar{x} \bar{y}$$

$$\sum (h k) = \sum (x_i y_i) - n \bar{x} \bar{y}$$

So

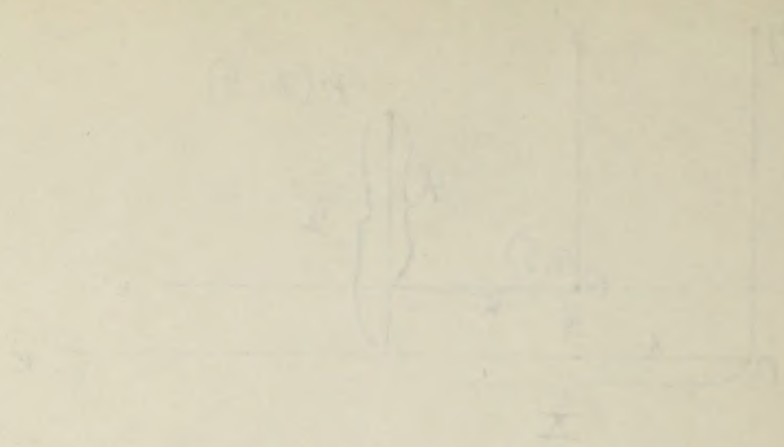
$$r = \frac{\sum h k}{N \sigma_x \sigma_y} = \frac{\sum (x y) - n \bar{x} \bar{y}}{n \sigma_x \sigma_y}$$

(5)

$$r = \frac{\frac{1}{N} \sum (x y) - \bar{x} \bar{y}}{\sigma_x \sigma_y}$$

Formula IV

Origin at an arbitrary point.



$$\text{let } h = x - \bar{x}$$

$$k = y - \bar{y}$$

In the formula  $r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$ ,  $x$  and  $y$  represent the deviations from the mean, and in the above  $h$  and  $k$  are the respective deviations, so we can

rewrite

$$r = \frac{\sum h k}{\sqrt{\sum h^2 \sum k^2}}$$

so if we let  $h = x - \bar{x}$  and  $k = y - \bar{y}$ , then

$\sum h k = \sum (x - \bar{x})(y - \bar{y})$

$\sum h^2 = \sum (x - \bar{x})^2$

$\sum k^2 = \sum (y - \bar{y})^2$

and so

$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$

which is the formula for the coefficient of correlation.

Example 1: Find the coefficient of correlation for the following data.

$x: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$

$y: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20$

Solution: First we find the mean of  $x$  and  $y$ .

|            |      |
|------------|------|
| $\sum x$   | 55   |
| $\sum y$   | 110  |
| $\sum x^2$ | 385  |
| $\sum y^2$ | 1540 |
| $\sum xy$  | 1100 |

(8)



Now for one more formula.

$$\sigma_x^2 = \frac{1}{N} \sum (x^2)$$

$$\sigma_x = \sqrt{\frac{1}{N} \sum (x^2)}$$

referred to the mean as origin.

Or if in the above diagram  $h = x - \bar{x}$

$$\sigma_h = \sqrt{\frac{1}{N} \sum (h^2)}$$

$$\begin{aligned} \sum (h^2) &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \\ &= (x_1^2 + x_2^2 + \dots + x_n^2) - 2\bar{x}(x_1 + x_2 + \dots + x_n) + n\bar{x}^2 \\ &= \sum (x_i^2) - 2\bar{x} \sum (x_i) + n\bar{x}^2 \\ &= \sum (x_i^2) - 2\bar{x} n\bar{x} + n\bar{x}^2 \end{aligned}$$

But

$$\sum (h^2) = \sum (x_i^2) - n\bar{x}^2$$

$$\therefore \sigma_h = \sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2}$$

Similarly

$$\sigma_y = \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}$$

$$\therefore r = \frac{\frac{1}{N} \sum (xy) - \bar{x} \bar{y}}{\sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2} \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}}$$

Formula V

Origin at an arbitrary point.

Summary of Formulas for r.

Where x and y are deviations from the true mean.

$$\text{I} \quad r = \frac{\frac{1}{N} \sum xy}{\sigma_x \sigma_y}$$

$$\text{II} \quad r = \frac{\sum xy}{\sqrt{\sum (x^2)} \sqrt{\sum (y^2)}}$$

Where x and y are deviations from an assumed mean.

$$\text{III} \quad r = \frac{\frac{1}{N} \sum (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}$$

$$\text{IV} \quad r = \frac{\frac{1}{N} \sum (xy) - \bar{x} \bar{y}}{\sigma_x \sigma_y}$$

$$\text{V} \quad r = \frac{\frac{1}{N} \sum (xy) - \bar{x} \bar{y}}{\sqrt{\frac{1}{N} \sum (x^2) - \bar{x}^2} \sqrt{\frac{1}{N} \sum (y^2) - \bar{y}^2}}$$

For one more formula.

referred to the mean as origin.

Or it is the same diagram

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

$$2(x - \bar{x})^2 = (x - \bar{x})^2 + (x - \bar{x})^2$$

Formula Y

Origin at an arbitrary point.

$$\frac{1}{N} \sum (x - \bar{x})^2 = \frac{1}{N} \sum (x - \bar{x})^2$$

Summary of formulas for

There x and y are deviations from the same origin.

$$I \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$II \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

There x and y are deviations from an assumed origin.

$$III \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$IV \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$V \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$VI \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$



Problem: Correlation<sup>of</sup> the Scores Received by 106 Reading School Pupils in the Henmon-Nelson and ~~The~~ Terman I. Q. Tests.

---

| Pupil | Henmon-Nelson | Terman | Pupil | Henmon-Nelson | Terman |
|-------|---------------|--------|-------|---------------|--------|
| 1     | 149           | 134    | 54    | 107           | 106    |
| 2     | 144           | 122    | 55    | 107           | 114    |
| 3     | 141           | 132    | 56    | 107           | 109    |
| 4     | 139           | 142    | 57    | 106           | 112    |
| 5     | 136           | 125    | 58    | 106           | 116    |
| 6     | 135           | 139    | 59    | 106           | 104    |
| 7     | 135           | 139    | 60    | 105           | 117    |
| 8     | 134           | 121    | 61    | 105           | 114    |
| 9     | 132           | 123    | 62    | 104           | 117    |
| 10    | 131           | 122    | 63    | 104           | 130    |
| 11    | 130           | 122    | 64    | 104           | 119    |
| 12    | 130           | 124    | 65    | 103           | 109    |
| 13    | 130           | 123    | 66    | 103           | 110    |
| 14    | 129           | 131    | 67    | 103           | 109    |
| 15    | 128           | 126    | 68    | 103           | 112    |
| 16    | 127           | 105    | 69    | 102           | 109    |
| 17    | 126           | 132    | 70    | 102           | 125    |
| 18    | 125           | 120    | 71    | 102           | 112    |
| 19    | 125           | 114    | 72    | 102           | 101    |
| 20    | 124           | 124    | 73    | 101           | 103    |
| 21    | 123           | 130    | 74    | 101           | 112    |
| 22    | 121           | 119    | 75    | 101           | 112    |
| 23    | 120           | 111    | 76    | 101           | 109    |
| 24    | 120           | 118    | 77    | 100           | 98     |
| 25    | 120           | 117    | 78    | 99            | 100    |
| 26    | 120           | 120    | 79    | 99            | 110    |
| 27    | 117           | 128    | 80    | 98            | 99     |
| 28    | 117           | 102    | 81    | 98            | 104    |
| 29    | 117           | 108    | 82    | 98            | 117    |
| 30    | 116           | 125    | 83    | 98            | 106    |
| 31    | 115           | 126    | 84    | 97            | 110    |
| 32    | 115           | 114    | 85    | 96            | 92     |
| 33    | 115           | 112    | 86    | 95            | 97     |
| 34    | 114           | 122    | 87    | 95            | 97     |
| 35    | 114           | 118    | 88    | 95            | 124    |
| 36    | 113           | 120    | 89    | 94            | 103    |
| 37    | 113           | 119    | 90    | 94            | 109    |
| 38    | 112           | 126    | 91    | 91            | 101    |
| 39    | 112           | 123    | 92    | 91            | 104    |
| 40    | 112           | 110    | 93    | 91            | 105    |
| 41    | 111           | 120    | 94    | 90            | 100    |
| 42    | 111           | 111    | 95    | 90            | 104    |
| 43    | 111           | 109    | 96    | 90            | 110    |
| 44    | 110           | 115    | 97    | 88            | 95     |
| 45    | 110           | 104    | 98    | 87            | 93     |
| 46    | 110           | 117    | 99    | 87            | 99     |
| 47    | 110           | 121    | 100   | 87            | 97     |
| 48    | 109           | 127    | 101   | 84            | 91     |
| 49    | 109           | 121    | 102   | 80            | 105    |
| 50    | 109           | 106    | 103   | 80            | 94     |
| 51    | 109           | 112    | 104   | 77            | 80     |
| 52    | 108           | 114    | 105   | 77            | 83     |
| 53    | 108           | 102    | 106   | 75            | 84     |





# Correlation Between Henman - Nelson 9.6. and Terman Group Test 9.6 Given to 106 Reading School Children.

| Nelson       | 75  | 80  | 85  | 90  | 95  | 100 | 105 | 110 | 115 | 120 | 125 | 130 | 135 | 140 | 145 |                   |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------------------|
| Terman       | 79  | 84  | 89  | 94  | 99  | 104 | 109 | 114 | 119 | 124 | 129 | 134 | 139 | 144 | 149 | $\Sigma xy$       |
|              | -6  | -5  | -4  | -3  | -2  | -1  | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |                   |
| 145-149      |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     | 7                 |
| 140-144      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 6                 |
| 135-139      |     |     |     |     |     |     |     |     |     |     |     |     | 2   |     |     | 5                 |
| 130-134      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 4                 |
| 125-129      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 3                 |
| 120-124      |     |     |     |     |     |     |     |     |     |     |     |     | 2   |     |     | 2                 |
| 115-119      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 110-114      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 105-109      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 100-104      |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 95-99        |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 90-94        |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 85-89        |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 80-84        |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| 75-79        |     |     |     |     |     |     |     |     |     |     |     |     | 1   |     |     | 1                 |
| $\Sigma x$   | 3   | 3   | 4   | 8   | 11  | 16  | 14  | 14  | 7   | 7   | 6   | 6   | 4   | 2   | 1   | $N=106$           |
| $\Sigma x^2$ | -6  | -5  | -4  | -3  | -2  | -1  | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | $\Sigma(x^2)=15$  |
| $\Sigma y$   | -18 | -15 | -16 | -24 | -22 | -16 | 0   | 14  | 14  | 21  | 24  | 30  | 24  | 14  | 8   | $\Sigma(y^2)=638$ |
| $\Sigma y^2$ | 108 | 75  | 64  | 72  | 44  | 16  | 0   | 14  | 28  | 63  | 96  | 150 | 144 | 98  | 64  | $\Sigma(xy)=638$  |

$$r = \frac{\frac{1}{N} \Sigma xy - \bar{x} \bar{y}}{\sqrt{\frac{1}{N} \Sigma x^2 - \bar{x}^2} \sqrt{\frac{1}{N} \Sigma y^2 - \bar{y}^2}}$$

$$\sigma_x = \sqrt{\frac{1}{N} \Sigma(x^2) - \bar{x}^2} = \sqrt{\frac{1036}{106} - 1.296} = 3.10$$

$$\sigma_y = \sqrt{\frac{1}{N} \Sigma(y^2) - \bar{y}^2} = \sqrt{\frac{638}{106} - 0.196} = 2.45$$

$$r = \frac{638}{106} - 0.36 \times 0.14 = \frac{3.10 \times 2.45}{7.60}$$

$$r = \frac{5.96 - 0.050}{7.60}$$

$$r = +0.77 +$$

The regression line.

$$X = \frac{\sigma_y}{\sigma_x} y = \frac{2.45}{3.1} y$$

$$X = \frac{77 \times 3.1}{2.45} y = \frac{77 \times 2.45}{3.1} y$$

$$X = 98 y$$

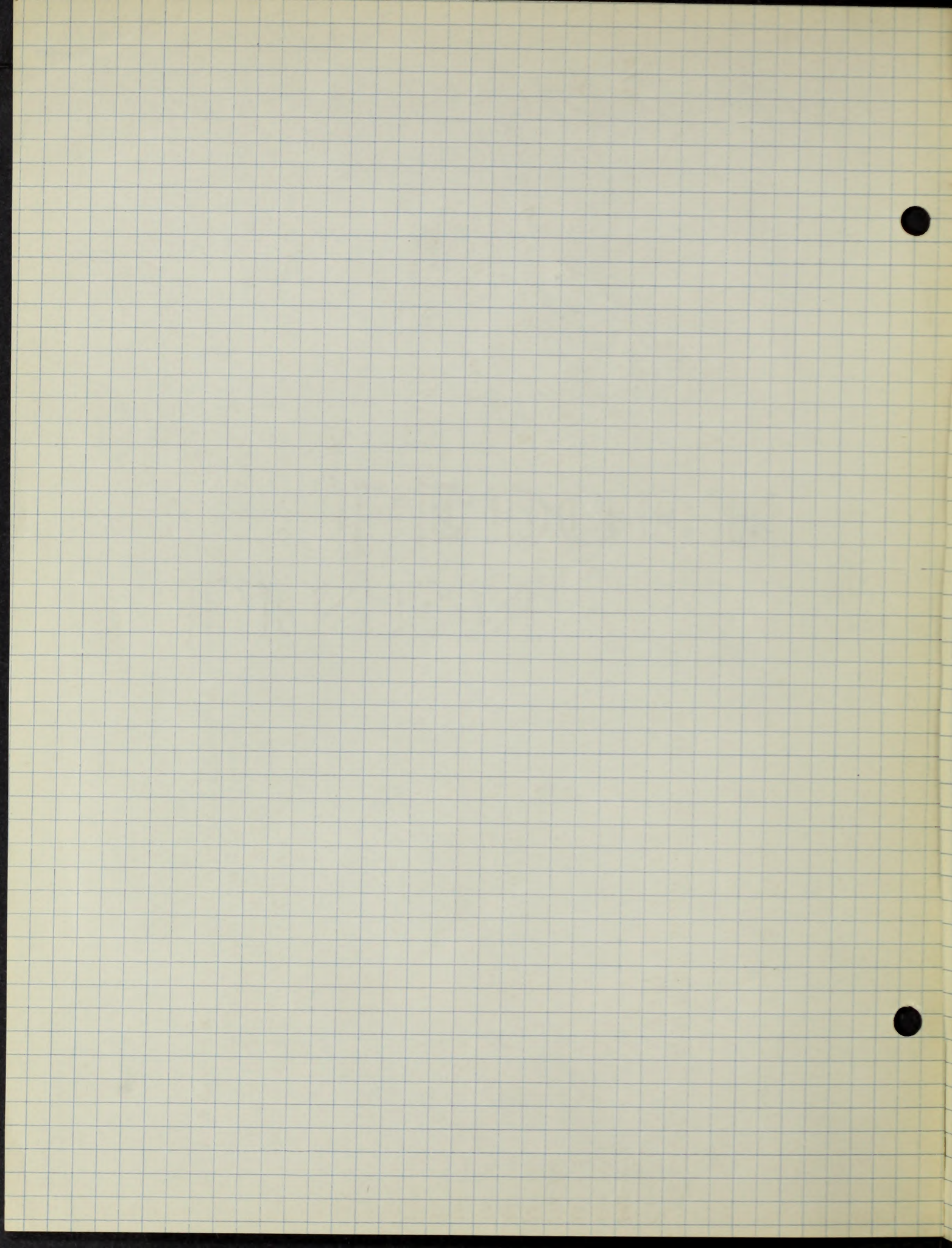
$$\bar{x} = \frac{38}{106} = 0.36$$

$$\bar{x}^2 = 1.296$$

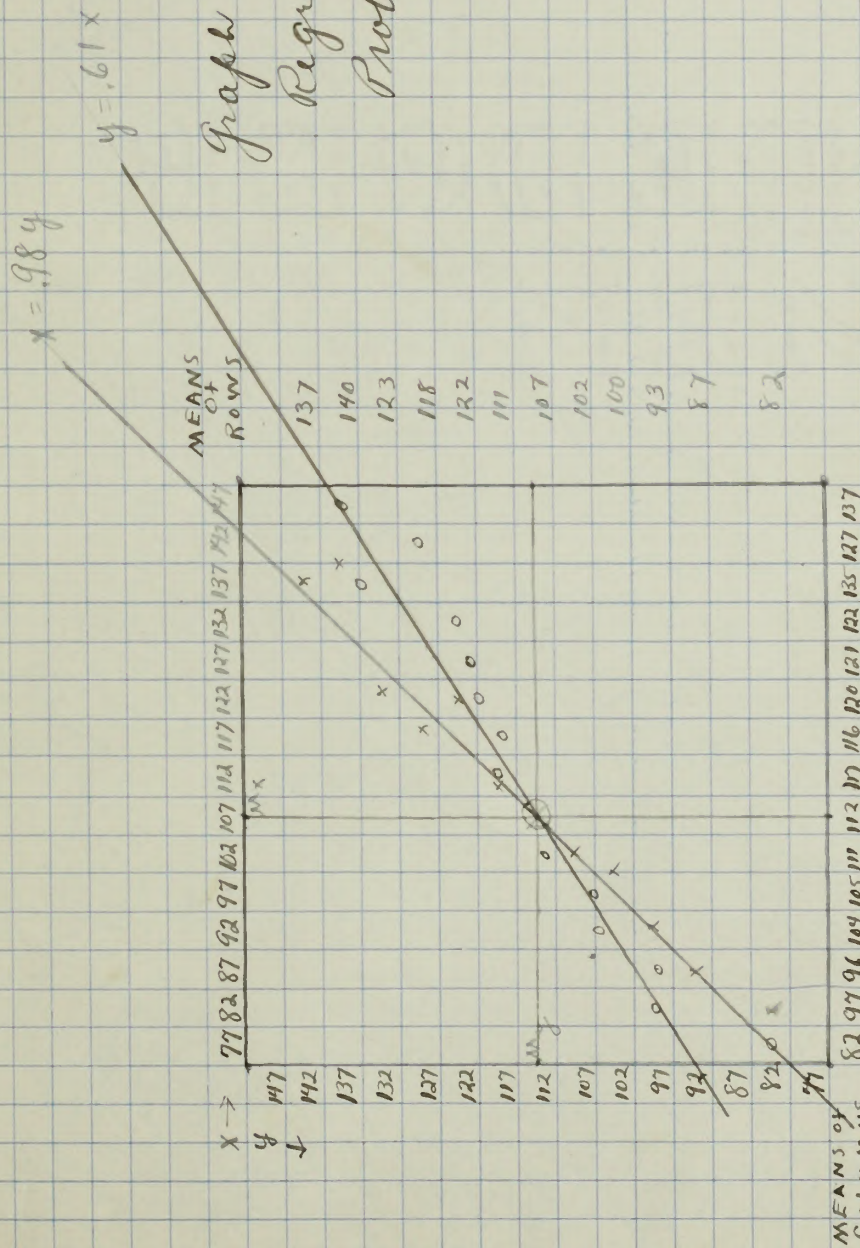
$$\bar{y} = \frac{15}{106} = 0.14$$

$$\bar{y}^2 = 0.196$$





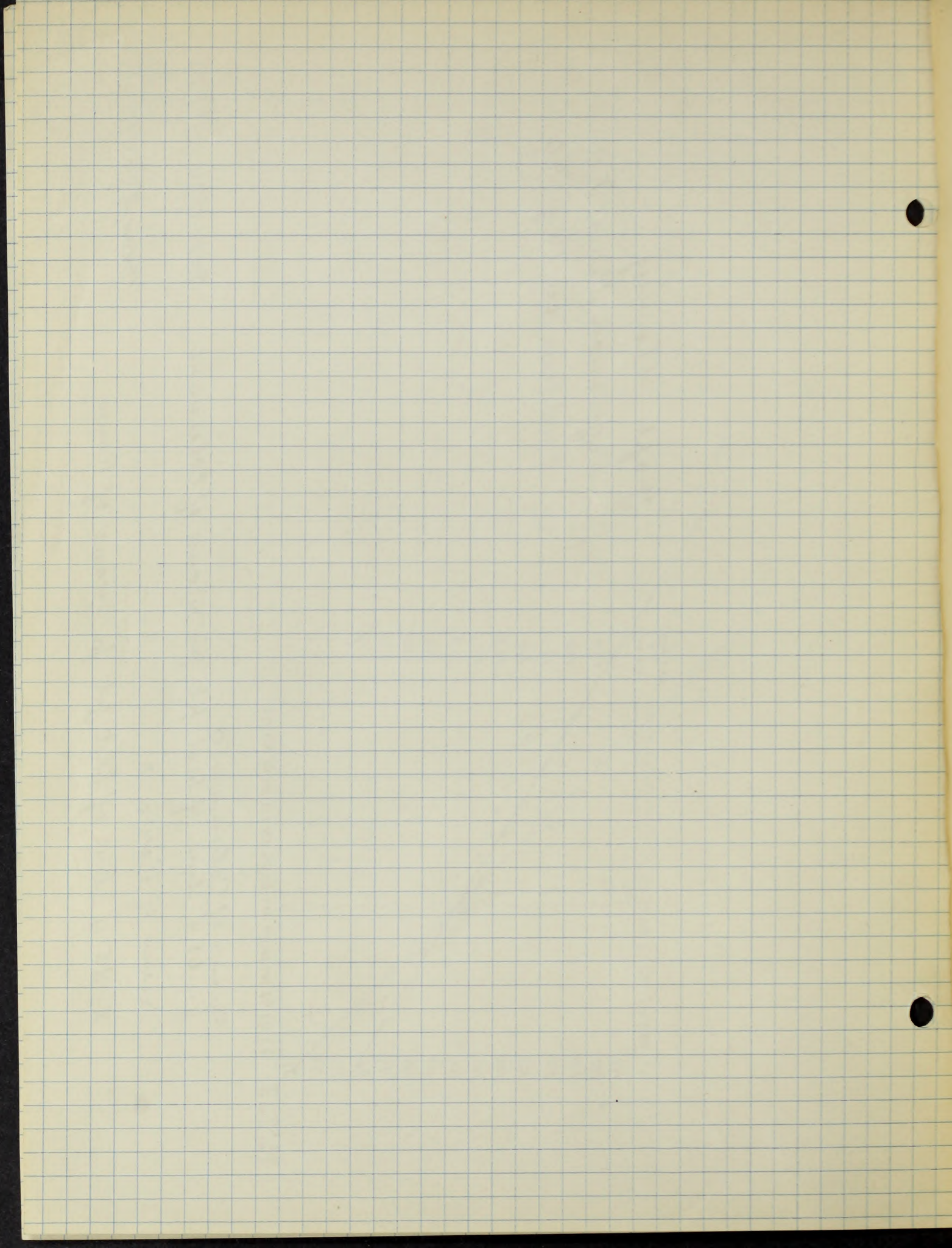




Graph showing  
Regression Lines For  
Problem on Page 46

Means of Columns  $\circ$   
 $y = .61x$  line of best fit for means of Columns  
 Means of Rows  $x$   
 $x = 98.4$  line of best fit for means of Rows  
 Mean of  $x's = 107.36$   
 Mean of  $y's = 112.13$







## Chapter VIII

Correlation From Ranks

## I. Introduction

When the data we are using expresses the measurements merely by the order or rank of the individual in the series, the product-moment formula for correlation is of no service in determining a measure of relationship. For example, consider the following table showing the ranks of ten students in an English and a history test.

|                      | A | B | C | D | E | F  | G  | H  | I  | J  |
|----------------------|---|---|---|---|---|----|----|----|----|----|
| Rank in English test | 1 | 2 | 3 | 4 | 5 | 6  | 7  | 8  | 9  | 10 |
| Rank in History test | 2 | 3 | 4 | 7 | 6 | 1  | 5  | 10 | 8  | 9  |
| Differences in Rank  | 1 | 1 | 1 | 3 | 1 | -5 | -2 | 2  | -1 | -1 |

We will try to show that if  $D$  is the difference in ranks of corresponding variables in the two series of  $N$  individuals then the correlation between the ranks is given by 
$$r = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

Now it can be seen easily that correlation between actually measured variables can be made to change without changing ranks. For example, consider these series:

|          |   |    |    |   |   |
|----------|---|----|----|---|---|
| Variates | x | -2 | -1 | 1 | 2 |
|          | y | -2 | -1 | 1 | 2 |
| Ranks    |   | 1  | 2  | 3 | 4 |
|          |   | 1  | 2  | 3 | 4 |

The correlation of the variables and the correlation of the ranks are perfect.

|          |   |    |      |     |   |
|----------|---|----|------|-----|---|
| Variates | x | -2 | -1.9 | 1.9 | 2 |
|          | y | -2 | -0.1 | 0.1 | 2 |
| Ranks    |   | 1  | 2    | 3   | 4 |
|          |   | 1  | 2    | 3   | 4 |

1. Introduction

When the data on two variables are arranged in a table, the order or rank of the individual in one series, the product-moment formula for correlation is of no service in determining the nature of relationship. For example, consider the following table showing the ranks of two students in mathematics and history tests.

|    | Mathematics | History |
|----|-------------|---------|
| 1  | 1           | 1       |
| 2  | 2           | 2       |
| 3  | 3           | 3       |
| 4  | 4           | 4       |
| 5  | 5           | 5       |
| 6  | 6           | 6       |
| 7  | 7           | 7       |
| 8  | 8           | 8       |
| 9  | 9           | 9       |
| 10 | 10          | 10      |

We will try to show that if  $r$  is the difference in ranks of corresponding variables in the two series of  $N$  individuals then the correlation between the ranks is given by  $r = 1 - \frac{1}{N}$ . It can be seen easily that correlation between ranks actually

measured variables can be made to change without changing ranks. For example, consider these variables:

|   |   |   |   |   |   |   |   |   |    |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Variables

|   |   |   |   |   |   |   |   |   |    |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

The correlation of the variables and the correlation of the ranks are

perfect.

|   |   |   |   |   |   |   |   |   |    |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |



Here the correlation of the ranks is still perfect but not so the correlation of the variates.

This would indicate that the value of  $\rho$  is not worth much by itself for interpretation and would show the necessity of connecting  $\rho$  with  $r$ , the coefficient of correlation of the variates. We will, therefore, try to show that under the assumption of a normal frequency distribution, and the assumption that grades may be replaced by ranks, the corresponding value of the correlation coefficient of the variates that correspond to the ranks is given by

$$r = 2 \sin \left( \frac{\pi}{6} \rho \right)$$

II. The Formula 
$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$

Reference: T.L.Kelley "Statistical Methods" P. 191-4

If  $x$  and  $y$  be the deviations from the mean of two variables to be correlated and if

$$\sigma_D^2 = \frac{\sum (x-y)^2}{N}$$

then 
$$\sigma_D^2 = \sigma_x^2 - 2r\sigma_x\sigma_y + \sigma_y^2$$

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_D^2}{2\sigma_x\sigma_y}$$

If  $\sigma_x = \sigma_y$ , then 
$$r = 1 - \frac{\sigma_D^2}{2\sigma^2}$$

Now if we are considering the coefficient of correlation between two series measured in rank only, each series contains  $N$  terms, the standard deviations and the means of each are equal respectively. The difference between the actual ranks of any one character would be equal to the differences of their deviations from the mean, so we may use the above formula where the coefficient of correlation for ranks is defined as  $\rho$ .

$$\therefore \rho = 1 - \frac{\sum D^2}{2N\sigma^2}$$

Here the correlation of the ranks is still perfect but not so the

correlation of the variables.

This result indicates that the value of  $\rho$  is not much affected by

for transformation and would show the necessity of considering  $\rho$  with  $r$ .

The coefficient of correlation of the variables is still, therefore,

very close to the correlation of a normal frequency distribution.

and the assumption that grades may be replaced by ranks, the corresponding

the value of the correlation coefficient of the variables that correspond

to the ranks is given by

$$\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}}$$

Reference: "The Theory of Statistics", p. 121-2

If  $x$  and  $y$  be the deviations from the mean of two variables to be

correlated and if

$$x = a_1 + a_2 + \dots + a_n$$

$$y = b_1 + b_2 + \dots + b_n$$

$$x^2 = a_1^2 + a_2^2 + \dots + a_n^2$$

$$y^2 = b_1^2 + b_2^2 + \dots + b_n^2$$

Then if we are considering the coefficient of correlation between two

series measured in rank only, each series contains  $n$  terms, the standard

deviations and the means of each are equal respectively. The difference

between the actual ranks of any one character would be equal to the dif-

ference of their deviations from the mean, and we may use the above

formula where the coefficient of correlation for ranks is defined as

$$\rho = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$



Now to show

$$\sigma^2 = \frac{N^2 - 1}{12}$$

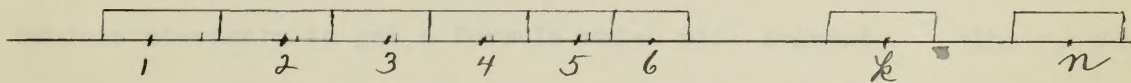
On Page 1 in the notes we show

$$S^2 = \sigma^2 + d^2$$

where  $S$  is the standard deviation about an arbitrary origin other than the mean. In this case let this origin be zero, then

$$d = \frac{1 + 2 + 3 + \dots + N}{N} = \frac{N + 1}{2}$$

$S^2$  is really the second moment of the ranks about zero, so we may determine  $S^2$  by first determining  $\overline{m}_2$ , the second moment about zero where the distribution consists of a frequency evenly spread over the class intervals, as shown below, instead of being concentrated at the midpoints as is the case where rank positions are used.



The frequency distribution drawn is represented by the line  $y = 1$  from  $x = \frac{1}{2}$  to  $x = N + \frac{1}{2}$ . The second moment of any one rank,  $k$ , from 0 is  $k^2$ , whereas the second moment of the distribution  $y = 1$  from  $k - \frac{1}{2}$  to  $k + \frac{1}{2}$  is

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} y x^2 dx = \left[ \frac{x^3}{3} \right]_{k-\frac{1}{2}}^{k+\frac{1}{2}} = k^2 + \frac{1}{12}$$

The second moment of the frequency  $y = 1$  corresponding to the  $k^{\text{th}}$  rank,

$\frac{1}{N}$  of the frequency, is  $\frac{1}{12}$  too large, as is true for every rank; hence the second moment of the equation  $y = 1$  from  $x = \frac{1}{2}$  to  $x = N + \frac{1}{2}$  will be larger than the desired second moment by  $\frac{1}{N} \left( \frac{N}{12} \right)$  or  $\frac{1}{12}$ . That is  $\overline{m}_2 = S^2 + \frac{1}{12}$

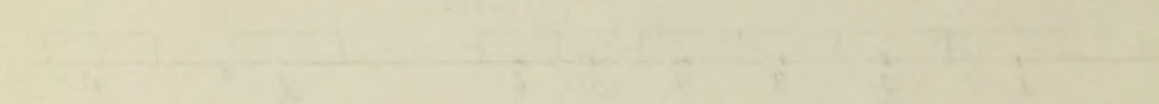
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

... is the standard deviation about an arbitrary origin other than the mean. In this case let this origin be zero, then

$$\mu = \frac{\sum x f(x)}{\sum f(x)}$$

... results the second moment of the curve about zero, so we may determine  $\sigma^2$  by first determining  $\mu^2$ , the second moment about zero

... the frequency curve consists of a frequency curve spread over the whole interval, as shown here, instead of being concentrated at one point as in the case where each position is fixed.



The frequency distribution shown is represented by the line  $y = f(x)$  and the second moment of the distribution is  $\mu^2$ . The second moment of the distribution is  $\mu^2$  from

$$\mu^2 = \frac{\sum x^2 f(x)}{\sum f(x)}$$

The second moment of the frequency curve is  $\mu^2$  corresponding to the  $x^2$  term,  $\mu^2$  of the frequency, is  $\mu^2$  and hence, as is true for every curve, hence the second moment of the function  $y = f(x)$  is  $\mu^2$  and  $\mu^2$  will be

hence the second moment of the function  $y = f(x)$  is  $\mu^2$  and  $\mu^2$  will be



$$\overline{m}_2 = \frac{1}{N} \int_{\frac{1}{2}}^{N+\frac{1}{2}} y x^2 dy = \frac{4N^2 + 6N + 3}{12}$$

$$S^2 = \frac{4N^2 + 6N + 2}{12}$$

$$\begin{aligned} \therefore \sigma^2 &= S^2 - d^2 \\ &= \frac{4N^2 + 6N + 2}{12} - \left(\frac{N+1}{2}\right)^2 \end{aligned}$$

$$\sigma^2 = \frac{N^2 - 1}{12}$$

$$\therefore \rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

III. The Formula  $r = 2 \sin\left(\frac{\pi}{6} \rho\right)$

Reference: Karl Pearson "On Further Methods of Determining Correlation" Drapers Company Research Memoirs Biometric Series IV.

Having found a formula for  $\rho$  for the correlation of ranks, it now becomes necessary to get a formula which will connect  $\rho$  with  $r$ , the coefficient of correlation for the variates. In the reference above Pearson develops such a formula

$$r = 2 \sin\left(\frac{\pi}{6} \rho\right)$$

where, however,  $\rho$  is the correlation for grades and not ranks. The rank is the actual position in order of an individual and is assumed to be at the midpoint of the class interval, hence if the rank is  $k$ , there are  $k - \frac{1}{2}$  individuals above that particular one in the series. Thus the grade of this particular individual would be  $k - \frac{1}{2}$ , the actual number above it in the series. Ranks form a discontinuous series with an interval of 1 while grades form a continuous series. The formula above may be used with ranks on the basis of two assumptions:

- (1) The series we are dealing with follow the normal law.
- (2) Grades of an individual may be replaced by ranks.

If we consider a series of  $N$  terms and each of these has a value in

$u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$   
 $u_t = \frac{1}{2} (u_{t-1} + u_{t+1})$

Theorem 1. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Having found a formula for  $v_t$  for the case of a constant  $u_t$ , it is not difficult to find a formula for  $v_t$  for the case of a linear  $u_t$ . In the latter case  $u_t = at + b$ , where  $a$  and  $b$  are constants. Then  $v_t = \frac{1}{2} (a(t-1) + b + a(t+1) + b) = at + b$ .

$$v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$$

Theorem 2. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 3. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 4. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 5. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 6. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 7. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 8. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 9. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

Theorem 10. Let  $u_t$  be a sequence of real numbers. Then the sequence  $v_t$  defined by  $v_t = \frac{1}{2} (u_{t-1} + u_{t+1})$  is also a sequence of real numbers.

If we consider a series of  $n$  terms and each of these has a value in



an x and one in a y frequency and we wish to find the coefficient of correlation, we may let  $m_1, m_2$  be the means;  $\sigma_1, \sigma_2$  the standard deviations;  $r$  be the correlation; and  $x$  and  $y$  the deviations from the means respectively. Then Pearson defines

$$g_1 = \frac{1}{2} N + \frac{N}{2\pi\sigma_1} \int_0^x e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}} dx$$

$$g_2 = \frac{1}{2} N + \frac{N}{2\pi\sigma_2} \int_0^y e^{-\frac{1}{2} \frac{y^2}{\sigma_2^2}} dy$$

where  $g_1$  and  $g_2$  are the x and y grades of the individuals in the series and  $\frac{1}{2} N$  is the mean of each. Since  $g_1$  and  $g_2$  are functions of  $x$  and  $y$ , correlation between  $g_1$  and  $g_2$  determines the correlation between  $x$  and  $y$  and vice versa.

Now if  $i_1 = g_1 - \bar{g}_1$ ;  $i_2 = g_2 - \bar{g}_2$ ;  
and

$$z = \frac{1}{2\pi\sigma_1\sigma_2} \cdot \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left( \frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right)}$$

then the product-moment of the grades is

$$(1) \quad P_{g_1, g_2} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 z dx dy$$

$$(2) \quad \frac{d P_{g_1, g_2}}{dr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{dz}{dr} dx dy$$

Pearson has shown in "Philosophical Transactions" Vol. 195A Page 25 that  $\frac{dz}{dr} = \sigma_1 \sigma_2 \frac{d^2 z}{dx dy}$ . The proof of this required the

definitions and notations for multiple correlation, so it has been assumed in this paper.

$$(3) \quad \frac{d P_{g_1, g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy$$

(4) Integrating twice by parts. (See Notes, Page 2)

$$\frac{d P_{g_1, g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$





(5) Substituting for  $\frac{dx}{dy}$  and  $\frac{dy}{dx}$  their values and letting

$$x = x'\sigma_1, y = y'\sigma_2 \quad (\text{See Notes, Page 3})$$

$$\begin{aligned} \frac{d \rho_{g, g_2}}{dr} &= \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)} \{ (2-r^2)x'^2 - 2rx'y' + (2-r^2)y'^2 \}} dx' dy' \\ &= \frac{N^3}{4\pi^2 \sqrt{1-r^2}} \cdot \frac{1}{\sqrt{\left(\frac{2-r^2}{1-r^2}\right)^2 - \frac{r^2}{(1-r^2)^2}}} \quad (\text{See Notes, Page 4}) \\ &= \frac{N^3}{2\pi \sqrt{4-r^2}} \end{aligned}$$

(6) Defining  $\rho = \frac{\rho_{g, g_2}}{N \sigma_{g_1} \sigma_{g_2}}$ ,  $\left\{ \begin{array}{l} \text{to correspond to the product-} \\ \text{moment formula for correlation} \end{array} \right.$

$$\frac{d\rho}{dr} = \frac{1}{N \sigma_{g_1} \sigma_{g_2}} \frac{d \rho_{g, g_2}}{dr}$$

$$\frac{d\rho}{dr} = \frac{6}{\pi^2} \cdot \frac{1}{\sqrt{4-r^2}}$$

(7)  $\rho = \frac{6}{\pi^2} \sin^{-1} \frac{1}{2} r + \text{a constant.}$

But since  $r$  is the coefficient of correlation between  $x$  and  $y$  and  $\rho$  is the coefficient of correlation between  $g_1$  and  $g_2$ ,  $\rho$  is zero when  $r$  is zero; therefore the constant above is zero.

$$(8) \therefore r = 2 \sin \left( \frac{\pi}{6} \rho \right)$$

(1) The correlation coefficient for the two variables is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \frac{100 - 10 \cdot 10}{\sqrt{100 \cdot 100}} = \frac{0}{100} = 0$$

(2) The correlation coefficient for the two variables is

(3) The correlation coefficient for the two variables is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \frac{100 - 10 \cdot 10}{\sqrt{100 \cdot 100}} = \frac{0}{100} = 0$$

(4) The correlation coefficient for the two variables is

(5) The correlation coefficient for the two variables is

(6) The correlation coefficient for the two variables is

(7) The correlation coefficient for the two variables is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



IV Problem Showing Correlation From Ranks Between Ten Students in a History and an English Examination.

| Student | Rank in English Test | Rank in History Test | Difference | Square of Difference |
|---------|----------------------|----------------------|------------|----------------------|
| A       | 1                    | 2                    | 1          | 1                    |
| B       | 2                    | 3                    | 1          | 1                    |
| C       | 3                    | 4                    | 1          | 1                    |
| D       | 4                    | 7                    | 3          | 9                    |
| E       | 5                    | 6                    | 1          | 1                    |
| F       | 6                    | 1                    | -5         | 25                   |
| G       | 7                    | 5                    | -2         | 4                    |
| H       | 8                    | 10                   | 2          | 4                    |
| I       | 9                    | 8                    | -1         | 1                    |
| J       | 10                   | 9                    | -1         | $\frac{1}{48}$       |

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$

$$\sum D^2 = 48, \quad 6 \sum D^2 = 288$$

$$N = 10, \quad N(N^2-1) = 990$$

$$\rho = 1 - \frac{288}{990} = 1 - 0.291 = 0.709$$

$$r = 2 \sin\left(\frac{\pi}{6} \rho\right) = 2 \sin 21.27^\circ = \underline{\underline{0.725}}$$

12. Problem: Showing Correlation from Scores Between Two Examinations in a History and an English Examination.

| Rank in English Test | Rank in History Test | Difference | Square of Difference |
|----------------------|----------------------|------------|----------------------|
| 1                    | 2                    | 1          | 1                    |
| 2                    | 3                    | 1          | 1                    |
| 3                    | 4                    | 1          | 1                    |
| 4                    | 7                    | 3          | 9                    |
| 5                    | 5                    | 1          | 1                    |
| 6                    | 1                    | -5         | 25                   |
| 7                    | 6                    | -1         | 1                    |
| 8                    | 10                   | 2          | 4                    |
| 9                    | 8                    | -1         | 1                    |
| 10                   | 9                    | -1         | 1                    |
|                      |                      |            | <hr/>                |
|                      |                      |            | 48                   |

$$G = 1 - \frac{440}{2500} = 0.824$$

$$S.D. = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{10} \cdot 48} = 2.196$$

$$G = 1 - \frac{440}{2500} = 0.824$$

$$r = \frac{G}{S.D.} = \frac{0.824}{2.196} = 0.375$$



## Chapter IX

Mean Square Contingency

## References:

- (1) "On the Theory of Contingency and Its Relation to Association and Normal Correlation" by Karl Pearson    Drapers' Company Research Memoirs, Biometric Series, I.
- (2) "Statistical Methods Applied to Education"    Harold O. Rugg    P. 299 et seq.
- (3) "Statistical Methods for Students in Education"    Holzinger    P. 273 et seq.
- (4) "Introduction to The Theory of Statistics"    Yule    64-67
- (5) "Introduction to Mathematical Statistics"    Carl J. West    Ch. 13

## I. Introduction

In the work with the coefficient of correlation, we were dealing with measured quantities, the statistics of variates. We now turn our attention to the relationship of traits which are not capable of quantitative measurement, the statistics of attributes.

A simple illustration will show the type of problem we are now to deal with. Suppose in a group of eighty-nine boys we wished to learn whether there was any association between their school work and their behavior and that these attributes could be tabulated as follows:

| School Work | Behavior |             |      |           |
|-------------|----------|-------------|------|-----------|
|             | Bad      | Troublesome | Good | Excellent |
| Good        | 3        | 9           | 12   | 14        |
| Medium      | 4        | 10          | 16   | 2         |
| Poor        | 10       | 2           | 7    | -         |

Clearly the product-moment method would not serve because we have no reasonable measurement for the various categories of behavior. We wish to find some method of measuring the amount of association which does not require us to determine scales for classifying the attributes.

THE UNIVERSITY OF CHICAGO  
LIBRARY  
1911

LIBRARY

THE UNIVERSITY OF CHICAGO  
LIBRARY  
1911

THE UNIVERSITY OF CHICAGO  
LIBRARY  
1911

THE UNIVERSITY OF CHICAGO  
LIBRARY  
1911



This method has been developed by Karl Pearson in his coefficient of mean square contingency.

## II. Contingency - Definition.

If we were considering the problem of the relationship of two attributes and classified them into a number of groups  $A_1, A_2, A_3$  - - - -  $A_s$  and  $B_1, B_2, B_3$  - - - -  $B_t$ , we would form a table containing  $s$  rows and  $t$  columns, or  $s \times t$  compartments with the total frequency distributed into sub-groups corresponding to these compartments.

|       | $B_1$ | $B_2$ | $B_3$ |  |  |  |  |  |  |  |  | $B_t$ |       |
|-------|-------|-------|-------|--|--|--|--|--|--|--|--|-------|-------|
| $A_1$ |       |       |       |  |  |  |  |  |  |  |  |       | $n_1$ |
| $A_2$ |       |       |       |  |  |  |  |  |  |  |  |       | $n_2$ |
|       |       |       |       |  |  |  |  |  |  |  |  |       |       |
| $A_s$ |       |       |       |  |  |  |  |  |  |  |  |       | $n_s$ |
|       | $m_1$ | $m_2$ | $m_3$ |  |  |  |  |  |  |  |  | $m_t$ | $N$   |

If the total frequency were  $N$  and if the numbers falling in the groups  $A_1, A_2$ , etc. were  $n_1, n_2$  - - -  $n_s$ , respectively (see table above) then the probability of one falling into one of these groups is  $\frac{n_1}{N}, \frac{n_2}{N}$  - - -  $\frac{n_s}{N}$  respectively. In like manner if the number falling in the groups  $B_1, B_2$  - - -  $B_t$  are  $m_1, m_2$  - - -  $m_t$  respectively, the probability of one falling in one of these groups will be  $\frac{m_1}{N}, \frac{m_2}{N}, \frac{m_3}{N}, \frac{m_4}{N},$  - - -  $\frac{m_t}{N}$ , respectively. Therefore, the number in the cell  $A_r B_c$  to be expected on the theory of independent probability is

$$N \cdot \frac{n_r}{N} \cdot \frac{m_c}{N} = \frac{n_r m_c}{N}$$

for the probability that a measure will fall in a row  $A_r$  is  $\frac{n_r}{N}$  and the probability that it will fall in a column  $B_c$  is  $\frac{m_c}{N}$ . Hence the probability that any one measure will fall in this row and column is

$\frac{n_r m_c}{N^2}$ , but there are  $N$  measures so the probability that any one will fall there is  $N \cdot \frac{n_r}{N} \cdot \frac{m_c}{N}$

This method has been developed by Karl Pearson in his coefficient of

contingency.

11. Contingency - Test.

If we want to test the hypothesis of the relationship of two

variables and classified them into a number of groups  $A_1, A_2, \dots, A_k$

and  $B_1, B_2, \dots, B_l$ , we would form a table as follows:

Let  $n_{ij}$  be the number of observations in the  $i$ th row and  $j$ th column, or a  $k \times l$  contingency table.

Let  $n_i$  be the total frequency in the  $i$ th row,  $n_j$  be the total frequency in the  $j$ th column, and  $n$  be the total frequency.

|          |          |          |          |          |          |
|----------|----------|----------|----------|----------|----------|
|          | $B_1$    | $B_2$    | $\dots$  | $B_l$    |          |
| $A_1$    | $n_{11}$ | $n_{12}$ | $\dots$  | $n_{1l}$ | $n_{1.}$ |
| $A_2$    | $n_{21}$ | $n_{22}$ | $\dots$  | $n_{2l}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $A_k$    | $n_{k1}$ | $n_{k2}$ | $\dots$  | $n_{kl}$ | $n_{k.}$ |
|          | $n_{.1}$ | $n_{.2}$ | $\dots$  | $n_{.l}$ | $n$      |

Let  $n_{i.}$  be the total frequency in the  $i$ th row,  $n_{.j}$  be the total frequency in the  $j$ th column, and  $n$  be the total frequency.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i = \frac{n_{i.}}{n}$  be the probability of an observation falling in the  $i$ th row,  $q_j = \frac{n_{.j}}{n}$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij} = \frac{n_{ij}}{n}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.

Let  $p_i$  be the probability of an observation falling in the  $i$ th row,  $q_j$  be the probability of an observation falling in the  $j$ th column, and  $p_{ij}$  be the probability of an observation falling in the  $i$ th row and  $j$ th column.



If the number actually observed in this cell is  $n_{rc}$  then

$$n_{rc} - \frac{n_r m_c}{N} \quad \text{measures the deviation from}$$

independent probability of the measure falling in the compartment

$A_r B_c$ .

Pearson points out that the total deviation of the whole system from independent probability must be some function of  $n_{rc} - \frac{n_r m_c}{N}$  for the whole table and he terms this total deviation from independent probability a measure of contingency. Therefore the greater the contingency, the greater must be the amount of correlation between the two attributes, for such a correlation is the measure of the degree of deviation from independence of occurrence. Pearson then points out that if we define

$$\lambda^2 = \sum \left\{ \frac{\left( n_{rc} - \frac{n_r m_c}{N} \right)^2}{\frac{n_r m_c}{N}} \right\}$$

we will have a function of  $n_{rc} - \frac{n_r m_c}{N}$  which will measure the degree of deviation of the series from independent probability and which will bring contingency into line with the customary notations of correlation. The formula used above is of the type developed by Elderton in "Frequency Curves and Correlation" on page 141 to measure the amount of agreement between two sets of figures. Here it is used to measure the amount of agreement between our observed data and the data of a table based on chance alone.

Definition: Mean Square Contingency.

Having defined  $\lambda^2$ , Pearson then defines  $\phi^2$ , the mean square contingency

$$\phi^2 = \frac{\lambda^2}{N}$$

The number actually observed is  $N$  and the number expected is  $N_0$ . The difference between the two is  $N - N_0$ .

Independent probability of the number falling in the compartment

is  $\frac{1}{N}$ .

It is pointed out that the total deviation of the whole system from the expected probability must be some function of  $N$ . For the whole system, the total deviation from the expected probability is a measure of non-independence. Therefore, the greater the deviation, the greater must be the amount of correlation between the two attributes. For each correlation is the measure of the degree of deviation from independence of occurrence. Therefore, the greater the deviation from independence of occurrence, the greater the correlation.

$$\left\{ \begin{array}{l} N - N_0 \\ N \end{array} \right\} \quad \left\{ \begin{array}{l} N - N_0 \\ N \end{array} \right\} \quad \left\{ \begin{array}{l} N - N_0 \\ N \end{array} \right\}$$

we will have a function of  $N$  which will measure the degree of deviation of the system from independent probability and which will indicate non-independence. This is the necessary condition for correlation. The formula used here is of the type developed by Pearson in "Biometrika" and "Biometrika Supplement" on page 141 to measure the amount of agreement between two sets of figures. It is used to measure the amount of agreement between our observed data and the data of a table based on chance alone.

Definition: When  $N$  is large,  $N - N_0$  is small.

It is defined that  $N$  is the number of observations,  $N_0$  is the number expected.

Correlation



### III. Development of the Formula for Mean Square Contingency.

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

Let  $x$  and  $y$  denote the deviations from their respective means of two attributes,  $\sigma_x, \sigma_y$  are the standard deviations and  $r$  is the correlation. Then if the correlation table can be approximately represented by the normal correlation surface,

$$Z_0 = \frac{N}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)}$$

represents the frequency with no correlation as previously discussed

and

$$Z = \frac{N}{2\pi\sqrt{1-r^2}\sigma_x\sigma_y} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rx}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}$$

represents the frequency with which we are dealing; i.e. the frequency of the observed data.

Since  $\phi^2 = \frac{\lambda^2}{N}$  and

$$\lambda^2 = \sum \left\{ \frac{\left(n_{rc} - \frac{n_r m_c}{N}\right)^2}{\frac{n_r m_c}{N}} \right\}$$

then

$$\phi^2 = \frac{1}{N} \sum \left\{ \frac{\left(n_{rc} - \frac{n_r m_c}{N}\right)^2}{\frac{n_r m_c}{N}} \right\}$$

Therefore, if we sum over the entire table this reduces to

$$(1) \quad \phi^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{(Z - Z_0)^2}{N Z_0} dx dy$$

$$(2) \quad \text{Substituting for } Z \text{ and } Z_0 \text{ and let } x' = \frac{x}{\sigma_x}, y' = \frac{y}{\sigma_y}$$

$$\begin{aligned} \phi^2 = & \frac{1}{2\pi} \left\{ \frac{1}{1-r^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x'^2 \frac{1+r^2}{1-r^2} - \frac{2rx'y'}{1-r^2} + y'^2 \frac{1+r^2}{1-r^2}\right)} dx' dy' \right. \\ & - \frac{2}{\sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left\{x'^2 \frac{1}{1-r^2} - \frac{2rx'y'}{1-r^2} + y'^2 \frac{1}{1-r^2}\right\}} dx' dy' \\ & \left. + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x'^2 + y'^2)} dx' dy' \right\} \end{aligned}$$

Let  $x$  and  $y$  denote the deviations from their respective means of two attributes.  $\sigma_x$  and  $\sigma_y$  are the standard deviations and  $r$  is the correlation. Then if the correlation table can be approximately repre-

$$r = \frac{\sum xy}{n \sigma_x \sigma_y}$$

represents the frequency with no correlation as previously discussed

$$\frac{1}{n} \sum \left( \frac{x}{\sigma_x} - r \frac{y}{\sigma_y} \right)^2 = \frac{1}{n} \sum \frac{x^2}{\sigma_x^2} - 2r \frac{1}{n} \sum \frac{xy}{\sigma_x \sigma_y} + r^2 \frac{1}{n} \sum \frac{y^2}{\sigma_y^2}$$

Therefore, if we sum over the entire table the relation for

$$(1) \quad \frac{1}{n} \sum \left( \frac{x}{\sigma_x} - r \frac{y}{\sigma_y} \right)^2 = 1 - r^2$$

$$(2) \quad \text{Substituting for } x \text{ and } y \text{ and for } \sigma_x \text{ and } \sigma_y$$

$$\frac{1}{n} \sum \left( \frac{x}{\sigma_x} - r \frac{y}{\sigma_y} \right)^2 = 1 - r^2$$



$$(3) \quad \phi^2 = \frac{1}{1-r^2} \cdot \frac{1}{\sqrt{\left(\frac{1+r^2}{1-r^2}\right)^2 - \frac{4r^2}{(1-r^2)^2}}} - \frac{2}{\sqrt{1-r^2}} \cdot \frac{1}{\sqrt{\left(\frac{1}{(1-r^2)^2} - \frac{r^2}{(1-r^2)^2}\right)}} + 1$$

This follows from the fact that if  $ac > b^2$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 - 2bx + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac - b^2}}$$

See Notes, Page 4

(4) Simplifying

$$\phi^2 = \frac{1}{1-r^2} - 2 + 1$$

$$(5) \quad r = \pm \sqrt{\frac{\phi^2}{1+\phi^2}}$$

Some important conclusions can be drawn from this. Elderton P.148.

"(1) It shows clearly that  $r$  must lie between  $-1$  and  $1$ .

(2) Since the value of  $\phi^2$  will not be affected by the order of rows (or columns), it will be seen that it is permissible to interchange them, provided, of course, the whole column (or row) be moved at once.

(3) The proof shows that  $r$  will not necessarily be obtained exactly if a very small number of groups is used, because by using the integral calculus an infinite number of groups was assumed.

(4) We also assumed, however, that we were dealing with a perfectly smooth series; but since  $\chi^2$  is a measure of goodness of fit between the correlation and non-correlation figures, a very large number of groups gives undue prominence to chance deviation, due to the use of random sampling, and the value of  $r$  found from that of  $\phi^2$  may differ considerably from the value reached by the  $xy$  moment. Too fine a grouping may give a less accurate result than

Some important considerations are as follows (Robinson, 1942):

- (1) It is shown clearly that a great deal of information is lost when the value of  $\rho$  is assumed to be zero (or constant). It will be seen that it is impossible to recover the original values of  $\rho$  from the observed values of  $r$  and  $s$ .
- (2) The above shows that  $r$  will not necessarily be obtained exactly if a very small number of groups is used. However, by using the integral calculus an infinite number of groups can be assumed.
- (3) It is also assumed, however, that an increasing number of groups will lead to a better approximation to the true value of  $\rho$ . This is a question of goodness of fit between the observed and non-correlation lines, a very large number of groups gives values much nearer to the true value than a small number of groups. The value of  $\rho$  found from the data is also affected considerably by the value of  $r$  and  $s$  used.

Consequently, the above findings give a less accurate result than



a less fine one."

Pearson further points out that since  $\phi^2$  is a measure of deviation of the series from independent probability and therefore of the amount of association or correlation between the attributes involved, any function of this expression is also a proper measure. Therefore, in order to bring the coefficient of contingency into line with the notations used in the coefficient of correlation, he defines the coefficient of mean square contingency

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

- V. Necessity of Limiting the Use of the Coefficient of Contingency to 5 x 5 fold or Finer Classifications. Yule. P. 65-6.

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}} = \sqrt{\frac{\lambda^2}{N + \lambda^2}}$$

Yule shows that coefficients when "calculated on different systems of classification are not comparable with each other. It is clearly desirable, for practical purposes, that two coefficients calculated from the same data, classified in two different ways, should be, at least approximately, identical. With the present coefficient this is not the case: if certain data be classified in, say (1) 6 x 6 fold, (2) 3 x 3 fold form, the coefficient in the latter form tends to be the least. The greatest possible value is, in fact, only unity if the number of classes be infinitely great; for any finite number of classes the limiting value of C is the smaller, the smaller the number of classes." Yule "Introduction to Theory of Statistics" P. 65.

The proof of this statement follows:

...the series from independent probability and therefore of the amount  
of association or correlation between the attributes involved, any factor  
of this expression is also a proper measure. Therefore, in order to find  
the coefficient of contingency into line with the notations used in the  
coefficient of correlation, we define the coefficient of contingency as

$$C = \frac{1}{\sqrt{1 + \frac{1}{\lambda^2}}}$$

$$C = \frac{1}{\sqrt{1 + \frac{1}{\lambda^2}}}$$

1. Necessity of finding the law of the coefficient of contingency as  
a  $\chi^2$  test or other classification. Yule, p. 25-26.

$$C = \frac{1}{\sqrt{1 + \frac{1}{\lambda^2}}}$$

This shows that coefficients when calculated on different systems  
of classification are not comparable with each other. It is clearly de-  
sirable, for practical purposes, that the coefficients calculated from  
the same data, classified in two different ways, should be at least  
approximately identical. With the present coefficient this is not the  
case: it varies with the way in which the data are classified in, say (1) a  $\chi^2$  test, (2) a  $\chi^2$  test.  
To find, then, the coefficient in the latter form tends to be the least.  
The greatest possible value is, in fact, only unity if the number of  
classes be infinitely great; for any finite number of classes the  
limiting value of  $C$  is the smaller, the smaller the number of classes.

This is a contradiction to the theory of statistics, p. 25.

The word of this statement follows:



$$(1) \quad \lambda^2 = \sum \left\{ \frac{\left( n_{rc} - \frac{n_r m_c}{N} \right)^2}{\frac{n_r m_c}{N}} \right\}$$

$$(2) \quad \lambda^2 = \sum \left\{ \frac{(n_{rc})^2}{\frac{n_r m_c}{N}} \right\} - 2 \sum n_{rc} + \sum \frac{n_r m_c}{N}$$

$$(3) \quad \text{Let } \sum \left\{ \frac{(n_{rc})^2}{\frac{n_r m_c}{N}} \right\} = S$$

$$\text{Then } \lambda^2 = S - 2N + N = S - N$$

$$(4) \quad \therefore C = \sqrt{\frac{S - N}{S}}$$

Now suppose we are to deal with a  $t \times t$  fold classification in which  $n_r = m_r$  for all values of  $r$ ; and suppose, further, that the association between the two attributes is perfect so  $n_r m_r = n_r = m_r$  for all values of  $r$ , and the frequencies in the remaining cells are zero. The frequency is then concentrated in the diagonal compartments of the table. If we interpret our notation in the light of this hypothesis, we have:

$$n_{rc} = n_r = m_c \quad \text{or} \quad (n_{rc})^2 = n_r m_c$$

$$\therefore S = \sum \left\{ \frac{(n_{rc})^2}{\frac{n_r m_c}{N}} \right\} = \sum (N) = tN$$

So we may write

$$C = \sqrt{\frac{t-1}{t}}$$

This is the greatest value of  $C$  for a symmetrical  $t \times t$  - fold classification.

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (1)$$

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (2)$$

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (3)$$

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (4)$$

For example, we have  $\bar{x} = 1.5$  and  $\bar{y} = 2.5$  for the data in Table 1.

Table 1. Data for Example 1. The first column contains the values of  $x$ , and the second column contains the values of  $y$ .

Table 2. The values of the sample means  $\bar{x}$  and  $\bar{y}$  for the data in Table 1.

Table 3. The values of the sample variances  $s_x^2$  and  $s_y^2$  for the data in Table 1.

Table 4. The values of the sample covariance  $s_{xy}$  for the data in Table 1.

Table 5. The values of the sample correlation coefficient  $r$  for the data in Table 1.

Table 6. The values of the sample regression coefficients  $b_1$  and  $b_2$  for the data in Table 1.

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (5)$$

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (6)$$

Table 7. The values of the sample regression coefficients  $b_1$  and  $b_2$  for the data in Table 1.

$$\left\{ \begin{array}{l} \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \end{array} \right\} \quad (7)$$

Table 8. The values of the sample regression coefficients  $b_1$  and  $b_2$  for the data in Table 1.

Table 9. The values of the sample regression coefficients  $b_1$  and  $b_2$  for the data in Table 1.



Yule then shows for

|          |   |               |       |
|----------|---|---------------|-------|
| $t = 2$  | C | cannot exceed | 0.707 |
| $t = 3$  | " | "             | 0.816 |
| $t = 4$  | " | "             | 0.866 |
| $t = 5$  | " | "             | 0.894 |
| $t = 6$  | " | "             | 0.913 |
| $t = 7$  | " | "             | 0.926 |
| $t = 8$  | " | "             | 0.935 |
| $t = 9$  | " | "             | 0.943 |
| $t = 10$ | " | "             | 0.949 |

so that it is well to restrict the coefficient of contingency to 5  $\times$  5 or finer classifications where the maximum value of C will at least approximate unity.

#### VI. Problem.

The coefficient of mean square contingency may be used for data quantitatively measured as well as for that which is qualitative. It may be used where one series is quantitative and one qualitative. The following example is from Rugg, P. 305, and shows the steps in using such a coefficient.

Relation Between Mental Age and Pedagogical Age.

|                  |                     | Mental Age |    |    |    |    |    |    | Totals   |
|------------------|---------------------|------------|----|----|----|----|----|----|----------|
|                  |                     | 9          | 10 | 11 | 12 | 13 | 14 | 15 |          |
| Pedagogical Age. | Retarded 2 years    |            |    |    | 2  |    | 7  | 2  | 11       |
|                  | Retarded 1 year     |            | 1  |    | 4  | 9  | 3  | 1  | 18       |
|                  | Normal              |            |    | 3  | 8  | 4  | 1  |    | 16       |
|                  | Accelerated 1 year  |            | 5  | 10 | 6  | 2  |    |    | 23       |
|                  | Accelerated 2 years | 2          | 7  | 3  | 1  | 1  |    |    | 14       |
|                  | Totals              | 2          | 13 | 16 | 21 | 16 | 11 | 3  | $N = 82$ |

$$C = \sqrt{\frac{S - N}{S}}, \quad S = \sum \left\{ \frac{(n_{vc})^2}{\frac{n_v m_c}{N}} \right\}$$

|       |       |       |       |
|-------|-------|-------|-------|
| 0.707 | 0.707 | 0.707 | 0.707 |
| 0.812 | 0.812 | 0.812 | 0.812 |
| 0.923 | 0.923 | 0.923 | 0.923 |
| 0.989 | 0.989 | 0.989 | 0.989 |
| 0.993 | 0.993 | 0.993 | 0.993 |
| 0.993 | 0.993 | 0.993 | 0.993 |
| 0.993 | 0.993 | 0.993 | 0.993 |
| 0.993 | 0.993 | 0.993 | 0.993 |
| 0.993 | 0.993 | 0.993 | 0.993 |

as well as to verify the coefficient of correlation to be  
in these observations when the value of  $\rho$  is less  
than unity.

The coefficient of correlation is calculated by the  
method of least squares as well as by the method of  
moments. The two methods are equivalent and give the  
same result. The following example is from Table 1. The  
value of  $\rho$  is 0.993.

| Table 1. Data for       |  |  |  |  |  |  |  |  |  |
|-------------------------|--|--|--|--|--|--|--|--|--|
| Correlation Coefficient |  |  |  |  |  |  |  |  |  |
| Method of Least Squares |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |
| Method of Moments       |  |  |  |  |  |  |  |  |  |



Table giving

$$\frac{n_r m_c}{N}$$

| Pedagogical Age. | Mental Age          |      |      |      |      |      |      |      |
|------------------|---------------------|------|------|------|------|------|------|------|
|                  |                     | 9    | 10   | 11   | 12   | 13   | 14   | 15   |
|                  | Retarded 1 year     |      |      |      | 2.82 |      | 1.48 | 0.40 |
|                  | Retarded 2 years    |      | 2.85 |      | 4.61 | 3.51 | 2.42 | 0.66 |
|                  | Normal              |      |      | 3.12 | 4.10 | 3.12 | 2.15 |      |
|                  | Accelerated 1 year  |      | 3.65 | 4.49 | 5.89 | 4.49 |      |      |
|                  | Accelerated 2 years | 0.34 | 2.22 | 2.73 | 3.59 | 2.73 |      |      |

The 2.85 in circle above is arrived at by the following

$$n_r = 18, \quad m_c = 13, \quad N = 82$$

$$\frac{n_r m_c}{N} = \frac{13 \times 18}{82} = 2.85$$

Table Showing

$$\left\{ \frac{(n_r m_c)^2}{\frac{n_r m_c}{N}} \right\}$$

| Pedagogical Age. | Mental Age          |        |       |       |       |       |       |       |
|------------------|---------------------|--------|-------|-------|-------|-------|-------|-------|
|                  |                     | 9      | 10    | 11    | 12    | 13    | 14    | 15    |
|                  | Retarded 2 years    |        |       |       | 1.42  |       | 33.14 | 10    |
|                  | Retarded 1 year     |        | 0.351 |       | 3.471 | 23.08 | 3.727 | 1.515 |
|                  | Normal              |        |       | 2.88  | 15.61 | 5.13  | 0.465 |       |
|                  | Accelerated 1 year  |        | 6.85  | 22.27 | 6.11  | 0.891 |       |       |
|                  | Accelerated 2 years | 11.735 | 22.07 | 3.295 | 0.279 | 0.367 |       |       |

$$S = 174.656$$

$$N = 82$$

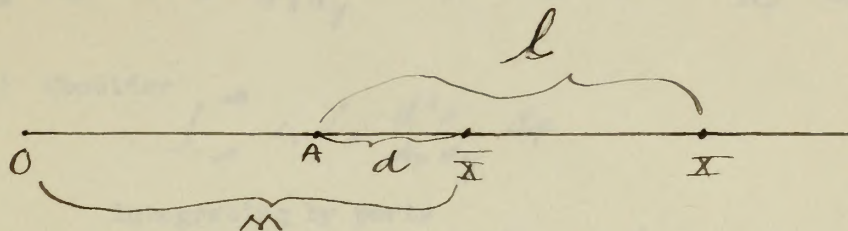
$$S-N = 92.656$$

$$C = \sqrt{\frac{92.656}{174.656}} = \sqrt{0.5305} = \underline{\underline{0.728}}$$





## I. Root Mean-Square Deviation



Definition of root mean-square

$$S^2 = \frac{1}{N} \sum (l^2)$$

or  $S^2$  is the average of the squares of the deviations about an arbitrary origin A.

$$(1) \quad l = \bar{X} - A$$

$$(2) \quad M - A = d$$

$$(3) \quad l = \bar{X} + d - M$$

$$(4) \quad l = \bar{X} - M + d$$

$$(5) \quad \text{Let } x = \bar{X} - M$$

$$(6) \quad \text{Then } l = x + d$$

$$(7) \quad l^2 = x^2 + 2xd + d^2$$

$$(8) \quad \sum (l^2) = \sum (x^2) + 2d \sum (x) + \sum (d^2)$$

$$(9) \quad \sum (x) \quad \text{the sum of the deviations about the mean and equals zero.}$$

$$(10) \quad \sum (l^2) = \sum (x^2) + \sum (d^2)$$

$$(11) \quad \frac{\sum (l^2)}{N} = \frac{\sum (x^2)}{N} = \frac{\sum (d^2)}{N}$$

$$(12) \quad S^2 = \sigma_x^2 + d^2$$

Root Mean-Square Deviation



Definition of root mean-square

$$\sigma = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

or  $\sigma$  is the square of the squares of the deviations about the

arithmetic mean.

$$(1) \quad C - X = 0$$

$$(2) \quad M - A = 0$$

$$(3) \quad X + 0 = M$$

$$(4) \quad X = M$$

$$(5) \quad X = M$$

$$(6) \quad X = M$$

$$(7) \quad X = M$$

$$(8) \quad X = M$$

$$(9) \quad X = M$$

the sum of the deviations about the mean and

results are:

$$(10) \quad \sum (x_i - \bar{x}) = 0$$

$$(11) \quad \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$(12) \quad \sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$



II. To show

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy.$$

(1) Consider

$$\int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx$$

Integrating by parts

$$= i_1(x) i_2(y) \left. \frac{dz}{dy} \right|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx$$

$$= - \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx$$

$$(2) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dx dy$$

$$(3) \text{ Now consider } \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} \frac{di_1}{dx} dy$$

$$= \frac{di_1}{dx} \int_{-\infty}^{\infty} i_2 \frac{dz}{dy} dy$$

Integrating by parts

$$= \frac{di_1}{dx} \left\{ \left[ i_2(y) z \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} z \frac{di_2}{dy} dy \right\}$$

$$= - \frac{di_1}{dx} \int_{-\infty}^{\infty} z \frac{di_2}{dy} dy$$

$$(4) \therefore \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} i_1 i_2 \frac{d^2 z}{dx dy} dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$

17. To show

$$\int_a^b \frac{d}{dx} \left( \frac{1}{p(x)} \right) dx = \frac{1}{p(b)} - \frac{1}{p(a)}$$

(1) Consider

$$\int_a^b \frac{d}{dx} \left( \frac{1}{p(x)} \right) dx$$

Integrating by parts

$$= \left[ \frac{1}{p(x)} \right]_a^b - \int_a^b \frac{1}{p(x)^2} \frac{dp}{dx} dx$$

$$= \frac{1}{p(b)} - \frac{1}{p(a)} - \int_a^b \frac{1}{p(x)^2} \frac{dp}{dx} dx$$

$$= \frac{1}{p(b)} - \frac{1}{p(a)} - \left[ \frac{1}{p(x)} \right]_a^b = \frac{1}{p(b)} - \frac{1}{p(a)} - \frac{1}{p(b)} + \frac{1}{p(a)} = 0$$

(2) Now consider

$$\int_a^b \frac{d}{dx} \left( \frac{1}{p(x)} \right) dx$$

$$= \left[ \frac{1}{p(x)} \right]_a^b - \int_a^b \frac{1}{p(x)^2} \frac{dp}{dx} dx$$

Integrating by parts

$$= \left[ \frac{1}{p(x)} \right]_a^b - \int_a^b \frac{1}{p(x)^2} \frac{dp}{dx} dx$$

$$= \frac{1}{p(b)} - \frac{1}{p(a)} - \int_a^b \frac{1}{p(x)^2} \frac{dp}{dx} dx$$

$$= \frac{1}{p(b)} - \frac{1}{p(a)} - \left[ \frac{1}{p(x)} \right]_a^b = \frac{1}{p(b)} - \frac{1}{p(a)} - \frac{1}{p(b)} + \frac{1}{p(a)} = 0$$



III To show  $\frac{d \rho_{g, g_2}}{dr} = \frac{N^3}{2\pi\sqrt{4-r^2}}$

$$(1) \quad \frac{d \rho_{g, g_2}}{dr} = \sigma_1 \sigma_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z \frac{di_1}{dx} \frac{di_2}{dy} dx dy$$

$$i_1 = \frac{N}{\sqrt{2\pi}\sigma_1} \int_0^x e^{-\frac{1}{2} \frac{x'^2}{\sigma_1^2}} dx' , \quad i_2 = \frac{N}{\sqrt{2\pi}\sigma_2} \int_0^y e^{-\frac{1}{2} \frac{y'^2}{\sigma_2^2}} dy'$$

$$\frac{di_1}{dx} = \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2} \frac{x^2}{\sigma_1^2}} , \quad \frac{di_2}{dy} = \frac{N}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2} \frac{y^2}{\sigma_2^2}}$$

$$\text{Let } x = x' \sigma_1 , \quad y = y' \sigma_2$$

$$\frac{di_1}{dx} = \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2(1-r^2)} (1-r^2) x'^2}$$

$$\frac{di_2}{dy} = \frac{N}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2(1-r^2)} (1-r^2) y'^2}$$

$$z = \frac{N}{2\pi\sigma_1\sigma_2} \cdot \frac{1}{\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} (x'^2 - 2rx'y' + y'^2)}$$

$$(2) \quad \frac{d \rho_{g, g_2}}{dr} = \frac{N^3}{4\pi^2\sqrt{1-r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)} ((2-r^2)x'^2 - 2rx'y' + (2-r^2)y'^2)} dx' dy'$$

$$(3) \quad \frac{d \rho_{g, g_2}}{dr} = \frac{2\pi}{\sqrt{\left(\frac{2-r^2}{1-r^2}\right)^2 - \frac{r^2}{(1-r^2)^2}}} \cdot \frac{N^3}{4\pi^2\sqrt{1-r^2}} = \frac{N^3}{2\pi\sqrt{4-r^2}}$$

$$\text{for } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 - 2bxy + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac-b^2}}$$

$$\text{if } ac > b^2$$

see next page.

THE

of the

of the

of the

of the

of the

of the

of the

of the



IV. To show

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2 - 2bxy + cy^2)} dx dy = \frac{2\pi}{\sqrt{ac - b^2}}$$

$$\text{if } ac > b^2$$

(1) The index may be written

$$-\frac{a}{2} \left(x - \frac{b}{a} y\right)^2 + \frac{y^2}{2a} (ac - b^2)$$

(2) Integrating with respect to  $x$ , since

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{k^2}} dx = k\sqrt{\pi}, \text{ we have}$$

$$\sqrt{\frac{2\pi}{a}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2a} (ac - b^2)} dy$$

(3) Now integrating with respect to  $y$ , since

$$ac - b^2 > 0$$

$$\text{we have } \sqrt{\frac{2\pi}{a}} \cdot \sqrt{\frac{2\pi a}{ac - b^2}} = \frac{2\pi}{\sqrt{ac - b^2}}$$

1. To show

$$f(x) = \frac{1}{2} (x^2 + 1) \quad \text{for } x \geq 0$$

$$f(x) = \frac{1}{2} (x^2 - 1) \quad \text{for } x < 0$$

(1) The label may be written

$$f(x) = \frac{1}{2} (x^2 + 1) \quad \text{for } x \geq 0$$

(2) Integration with respect to x, since

$$f(x) = \frac{1}{2} (x^2 + 1) \quad \text{for } x \geq 0$$

$$f(x) = \frac{1}{2} (x^2 - 1) \quad \text{for } x < 0$$

(3) Now integration with respect to x, since

$$f(x) = \frac{1}{2} (x^2 + 1) \quad \text{for } x \geq 0$$

$$f(x) = \frac{1}{2} (x^2 + 1) \quad \text{for } x \geq 0$$



### Conclusion

In this paper I have tried to show something of the development of the formulas for simple correlation of three important types; the coefficient of correlation for linear regression, the most important in its frequent use; the coefficient of correlation from ranks; and the coefficient of mean square contingency. The first of these to be used where the data are represented by numerical measures and the method taking full account of the value and position of every measure in the series, the second to be used where only the positions of the measures are given, and the third when the data are not in terms of numerical measures but in the form of attributes. I have been interested in these because I felt they were suitable formulas for work usually done in statistics in Education. A more complete discussion should, of course, contain something of the correlation ratio to be used where the regression is non-linear, a study of tests for linearity, and a study of probable errors. These topics would form a suitable study in themselves.

In developing the coefficient of correlation by the correlation surface method certain assumptions based on the theory of probability were made, and the equation of the normal curve was used without deriving it. These could have been given a sound mathematical basis but it seemed wise to limit the paper and give references for their derivation.

In the chapter on correlation from ranks, the assumption was made that grades could be replaced by ranks. Pearson makes this assumption in the reference cited in that chapter. The sound method would be, it seems, to use grades exclusively but the work involved would be extremely laborious and the results not sufficiently different to warrant such an effort.





## Bibliography

- Camp, B. H. "The Mathematical Part of Elementary Statistics." 1931  
Chapters VIII, IX and X.
- Chaddock, Robert Emmet. "Principles and Methods of Statistics." 1925  
Chapter XII
- Elderton, W. Palin. "Frequency Curves and Correlation" Part II. 1906
- Forsyth, C. H. "An Introduction to the Mathematical Analysis of  
Statistics." 1924. Chapter X
- Holzinger, Karl J. "Statistical Methods for Students in Education"  
1928. Chapter IX
- Jones, D. C. "A First Course in Statistics." 1924. Chapters X and XIX
- Kelley, Truman L. "Statistical Methods" 1923. Chapter VIII
- Odell, C. W. "Educational Statistics" 1925. Chapter V
- Pearson, Karl "On the Theory of Contingency and Its Relation to  
Association and Normal Correlation," Drapers Company Research  
Memoirs, Biometric Series I. 1904
- Pearson, Karl "On Further Methods of Correlation" Drapers Company  
Research Memoirs, Biometric Series IV. 1907
- Rietz, H. L. "Mathematical Statistics." 1927. The Carus Mathematical  
Monograph, Number Three. Chapter IV
- Rietz, H. L. and Crathorne, A. R. "Handbook of Mathematical Statistics"  
Edited by H. L. Rietz. 1924. Chapter VIII
- Rugg, Harold O. "Statistical Methods Applied to Education" 1917  
Chapter IX
- West, Carl J. "Introduction to Mathematical Statistics" 1918  
Chapters VII and IX
- Yule, G. Udny "An Introduction to the Theory of Statistics" 1917  
Chapters IX and XVI











BOSTON UNIVERSITY



1 1719 02551 7394



